St. Paul's University Repository

CORPORATE REPUTATION MANAGEMENT IN THE ERA OF ARTIFICIAL INTELLIGENCE AND FAKE NEWS: A CASE STUDY OF SAFARICOM PLC

ERIC KAMANGA

ST. PAUL'S UNIVERSITY

A DISSERTATION SUBMITTED TO THE SCHOOL OF COMMUNICATION AND COMPUTER STUDIES, ST. PAUL'S UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS IN COMMUNICATION STUDIES

OCTOBER, 2026

St. Paul's University Repository

DECLARATION

This dissertation is my original work and has not been presented for a degree in any other university. No part of this dissertation/thesis may be reproduced without the prior written permission of the author and/or St. Paul's University.

Signature:  Date: 21/10/2026

Eric Kamanga

Word Count: 24,024

Declaration by Supervisors

This dissertation/thesis has been submitted for examination with our approval as university supervisors.

Signature:  Date.:21/10/2026

Dr Ukaiko Ojiambo

First Supervisor

Signature:  Date:21/10/2026

Andrew Yotui

Second Supervisor

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

ACKNOWLEDGMENT

First and foremost, I want to thank the Almighty God for giving me the strength, wisdom, grace and perseverance to complete this dissertation. His guidance has been the foundation of this journey. I sincerely thank my supervisors, Dr. Ukaiko Ojiambo and Mr. Andrew Yotui, for their guidance, patience, constructive feedback and support throughout this research. I sincerely appreciate all the research participants and key informants who generously gave their time and shared valuable insights that contributed to this study. To my loving Mother, Naomi Kamanga, thank you for your love, prayers, sacrifices and unwavering belief in me. Your encouragement has been a constant source of strength, and I am eternally grateful for everything you have done for me. I am equally grateful to my entire family and friends for their encouragement, patience, prayers and support throughout this journey. Finally, special mention to my son, Donell Kamanga. May this achievement inspire you to pursue excellence, value education and always strive to become the best version of yourself. May you grow up knowing that with faith, discipline, hard work and determination, no dream is beyond reach. To everyone who contributed to this journey in one way or another, thank you sincerely. This achievement would not have been possible without you. God Bless!

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

ABSTRACT

Artificial intelligence (AI) has revolutionised the corporate misinformation space, allowing for the quick generation of synthetic fraud content, deepfakes, and faux announcements that can overwhelm organisations' detection and response capabilities. While both Situational Crisis Communication Theory (SCCT) and Image Repair Theory (IRT) provide the most popular models to guide the handling of reputational crises, both theories were developed before the advent of generative AI and have not been tested systematically against threats posed by AI, particularly in the mobile-money sector, where Safaricom PLC's mobile-money platform M-Pesa dominates the market in Kenya. This study explored typologies of AI-generated misinformation targeting Safaricom, the crisis communication counter-strategies it used to respond, their effectiveness in protecting its corporate reputation and implications for stakeholder trust, and the continued relevance of SCCT and IRT in this context. The study adopted a qualitative, interpretivist approach using a single instrumental case study design, and triangulated data from key informant interviews with 24 respondents (including Safaricom employees, external stakeholders, and experts), documentary analysis of 41 documents, and content analysis of 3,412 social media posts from 2021 to 2026, analysed thematically following the six-step approach by Braun and Clarke. The top four types of misinformation included fabricated mobile-money fraud alerts, fabricated announcements, impersonation accounts and deepfakes. Safaricom responded mostly reactively and denialist in 55% of severe incidents, in line with its internal response-time target. Social media sentiment improved significantly after the official response, but stakeholders perceived the response as not as rapid, especially regarding the longevity of results achieved through closed communication channels. The current practice is limited by the detection lag, misinformation formation in the closed channel and overemphasis on the measures of the process of assessment. The results of the study reveal that SCCT and IRT need theoretical extension to fully capture the dynamics of fast-paced misinformation channels, their speed, and scale in AI-exposed markets, and provide practical tools and strategies to improve corporate crisis communication in these markets.

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

TABLE OF CONTENTS

DECLARATION.....i
ACKNOWLEDGMENT.....ii
-CHAPTER ONE: INTRODUCTION.....1
1.0 Background to the Study.....1
1.1 Statement of the problem.....6
1.4 Purpose of the Study.....7
1.5 Objectives.....7
1.6 Research questions.....8
1.7 Justification of the study.....8
1.8 Significance of the study.....9
1.10 Scope of the study.....9
1.11 Limitations and Delimitations of the Study.....10
CHAPTER TWO: LITERATURE REVIEW.....12
2.0 Introduction.....12
2.1 General literature review.....12
2.2 Empirical Literature Review.....14
2.3 Theoretical framework.....22
2.5 Conceptual Framework.....25
2.4 Summary of Research Gap.....26
2.6 Summary of the chapter.....28
CHAPTER THREE: METHODOLOGY.....30
3.0 Introduction.....30
3.1 Research Philosophy.....30
3.2 Research Methodology.....31
3.3 Research Design.....32
3.4 Population.....32
3.5 Target Population.....33
3.6 Sampling Technique.....33
3.8 Data Collection.....35

ST. PAUL'S UNIVERSITY

3.9 Trustworthiness and Rigour....38
3.10 Data Analysis....38
3.11 Ethical Considerations....39
3.12 Chapter Summary....40
CHAPTER FOUR: DATA PRESENTATION, ANALYSIS, AND INTERPRETATION41
4.1 Introduction....41
4.2 Participants and Corpus....41
4.3 Presentation, Analysis, and Interpretation....43
CHAPTER FIVE: SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS....55
5.1 Introduction....55
5.2 Conclusions....57
5.3 Recommendations....58
5.4 Areas for Future Studies....59
REFERENCES....60
APPENDICES....66
Appendix 1: Interview Guide: Safaricom Communication Professionals....66
Appendix 2: Interview Guide: External Stakeholders....71
Appendix 3: Interview Guide: Expert Informants....75
Appendix 4: Documentary Analysis Framework....79
Appendix 5: Social Media Content Analysis Protocol....84
Appendix 6: ISERC Approval Letter....89
Appendix 7: NACOSTI Certificate....90

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

CHAPTER ONE: INTRODUCTION

1.0 Introduction

The chapter opens by placing the study in the context of the current landscape of misinformation and the role artificial intelligence plays in it, as well as the implications for corporate reputation management. It starts with an introduction to the study, tracing the migration of AI technologies and their impact on the information landscape in which corporations are acting and interacting with their constituencies. The chapter then restates the problem in order to define the specific void that this study fills regarding the use of existing crisis communication theory in the context of AI-generated threats. This is followed by the purpose of the study, the objectives that guided the study, and the research questions that were designed to operationalise the specific objectives. The chapter also presents the rationale and relevance of the study, justifying its significance at a particular stage in the Kenyan context and explaining who benefits from the study.

Lastly, the chapter establishes the scope in which the study was conducted, recognizes the limitations and delimitations established to control them, and summarises the findings and concludes the chapter before finally moving into the Literature Review in Chapter Two that immerses the reader in the literature review. Artificial Intelligence (AI) is commonly referred to as the fourth industrial revolution (Oosthuizen, 2022). This transformation extends beyond productivity; it also changes corporate communications and reputation risk. This turning point in the presence of AI technologies in the networked society has made information easier for organizations to share at a greater rate and breadth than ever before, and likewise for anyone with bad intentions to fabricate, disguise, and spread false narratives about these organizations In this context, the current research focused on Safaricom PLC, a leading mobile-money service provider in Kenya, in addressing this dynamic risk.

1.0 Background to the Study

Digital transformation has become one of the most important trends in the modern business environment. Recent developments such as artificial intelligence, data analytics, and digital communication have been disruptive and among the most innovative business developments of modern times (Machucho & Ortiz, 2025). These technological advancements

ST. PAUL'S UNIVERSITY

are shaping how businesses interact with stakeholders and manage information and business operations. Artificial intelligence remains one of the most disruptive technologies, enabling faster information processing, automated communication systems, and predictive analytics that often improve efficiency and responsiveness. Nonetheless, despite these benefits, the technology has also posed significant challenges, especially regarding misinformation and fake news, which pose a serious threat to the corporation's reputation. Currently, AI-powered tools can generate realistic fake content and automate the spread of misinformation, amplifying misleading narratives across digital platforms and making it difficult for organisations to control information about themselves.

Artificial intelligence tools, especially the large language models (LLMs) and generative AI systems, have become commonplace in modern society (Abolghasemi et al., 2025). These tools are widely accepted and used to generate highly realistic content, including deepfake videos, fabricated press releases, manipulated executive speeches, and falsified financial announcements (Ma et al., 2025; Machucho & Ortiz, 2025). These capabilities have significantly altered the misinformation landscape for corporations. A global study Arcuri et al. (2023) examining the role of fake news and deepfakes in the American and European Union stock markets reports that misinformation directly influenced stock prices and investor behaviour. In a New York Post article, Sorkin et al. (2023) it is reported that an AI-generated image triggered a stock sell-off in the United States, as misinformation rattled the market and sparked panic selling.

Studies on the COVID-19 pandemic further reveal that LLM-generated fake news triggered panic-buying behaviours in the United Kingdom, illustrating how AI-driven misinformation can produce real-world economic and social consequences (Ali et al., 2023). Additionally, deepfakes caused Arup, a UK-based company, to lose millions of pounds after a fake executive meeting (Milmo, 2024). Collectively, these incidents highlight the growing urgency of proactive crisis communication strategies in the current age of artificial intelligence. This trend has been further demonstrated in Canada's infamous "Buttergate," where the use of AI fuelled misinformation about product shortages. AI has played an important role in the spread of misinformation using deepfakes and micro-targeting (Music et al., 2022). Through various tactics, AI can exploit cognitive overload, spread fake news, and optimize engagement, often resulting in widespread misinformation.

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

Corporate reputation is critical in shaping the collective perception of a company and its attractiveness to stakeholders, thereby impacting business outcomes (Nuortimo et al., 2024). In recent times, corporations are facing new threats to their business reputation stemming from artificial intelligence (AI). Notably, there has been an increase in AI-generated misinformation, ranging from false announcements and screenshots to manipulated executive speeches and fake press releases, which can affect an organization's financial performance and result in public mistrust. In the United States and the European Union, fake news and deepfakes have resulted in turbulence in the stock market (Arcuri et al., 2023). These incidents highlight the growing importance of proactive crisis communication strategies, especially in the current age of artificial intelligence.

Crisis communication plays a critical role in managing corporate reputation during crises or reputational threats. According to Seeger (2022) the concept, it can be understood as the process of creating and exchanging messages between interdependent stakeholders in conditions of high uncertainty, threats to high-priority goals, and the need for immediate response stemming from a crisis. Moreover, other scholars have defined crisis communication as a strategic process through which organizations communicate with stakeholders before, during, and after a crisis to reduce damage and restore trust (Upadhyay & Upadhyay, 2023). Effective communication often involves timely responses, transparency, accuracy, consistency, and empathy, while also promoting the organization's monitoring of communication channels to support community engagement. Traditional crisis management had clearly defined structures that stakeholders could follow. The modern AI environment has significantly changed this landscape due to the scale, speed, and sophistication of misinformation, underscoring the need for more proactive communication approaches to manage reputational crises effectively.

Globally, corporations are experiencing a decline in public trust, expertise, and media. This has been further exacerbated by the emergence of AI-driven tools, which have emerged as a non-cyber-related security risk (Vlaeminck, 2023). In the Global North, corporations are experiencing an emerging reputation crisis driven by AI-generated, amplified fake news. Misinformation about product quality, financial stability, and leadership conduct continues to shape stakeholder behaviour and erode consumer trust in sectors ranging from pharmaceutical

ST. PAUL'S UNIVERSITY

to financial services (Ma et al., 2025). In the United States and the EU, fake news events rattled equity markets, with effects that persisted even after corrective measures were implemented, suggesting that reputational damage outlasts the initial misinformation incident. This phenomenon was also reported in the UK during the COVID-19 pandemic, where fake news triggered panic buying, prompting market adjustments. These international trends highlight a new era of corporate vulnerability, in which the capacity to respond effectively to AI-driven fake news has become the core of organisational competency.

Nonetheless, this is not only a problem of the northern world. In Asia, the problem of AI misinformation remains a challenge. Sonni (2025) in a study conducted in Indonesia highlights that the emergence of AI has led to widespread misinformation through automated content production, resulting in the large-scale dissemination of hate-laden messages. Secondly, the use of AI has enabled the personalisation of content, helping to create an echo chamber that reinforces polarisation. Lastly, the study reports that multimodal manipulation of content has helped in improving the credibility of misleading content. Similarly, Ibáñez et al. (2023) a study conducted in China highlights the rise and current spread of AI misinformation, especially in journalism, which has the potential to erode the credibility of media houses and the public's overall trust.

. In Africa, AI-driven disinformation and misinformation have been well documented, especially during conflicts and election periods. In Sudan, there has been an increase in AI-generated content, a phenomenon scholars have labelled the "Lias Dividend" (Mohamed, 2025). This often involves false claims that materials are inauthentic to divert attention from themselves, evade accountability, or undermine adversaries (Okolo, 2024). Moreover, in Nigeria, studies show that AI played a significant role in shaping electoral outcomes and the overall campaign in recent elections. One of the most popular videos during the campaign season was of Elon Musk endorsing a presidential candidate (Awojulugbe, 2022). Additionally, deepfakes of former and current US presidents endorsing different presidential candidates circulated. While these incidents offer valuable insights into the phenomenon under study, they focus on the political aspect and overlook the corporate context.

In Sub-Saharan Africa, and Kenya in particular, Kenya's rapidly expanding digital infrastructure, characterised by high mobile penetration, widespread internet access, and a highly active social media user base, has simultaneously facilitated greater organisational-

ST. PAUL'S UNIVERSITY

stakeholder connectivity and heightened exposure to misinformation risks (Mwakideu, 2022). Kamau and Shiundu (2024) observe that misinformation has become widespread on Kenya's social media platforms, particularly Facebook, WhatsApp, X (formerly Twitter), and Instagram, which also serve as primary sources of information for young Kenyans. The consequences for institutional credibility have been significant. Kenya's Independent Electoral and Boundaries Commission (IEBC), the constitutional body mandated to administer national elections, experienced extensive AI-assisted disinformation campaigns during the 2022 general elections, in which fabricated results, manipulated data, and coordinated bot activity were deployed to undermine public confidence in the electoral process (Miyandazi & Thuo, 2025; Kamau & Shiundu, 2024). AI-generated deepfake videos and audio recordings of senior political figures also circulated widely during this period, contributing to what studies describe as an "information disorder" environment in which citizens struggle to distinguish authentic from fabricated content (Miyandazi & Thuo, 2025). The IEBC case is instructive for Kenyan corporations because it illustrates the scale and sophistication of AI-enabled disinformation in Kenya, challenging the notion that private organizations, including telecommunications companies, are equally exposed.

Safaricom was established in 1997 as a 100% subsidiary of Telkom Kenya. Vodafone invested 40 % in the company in May 2000, and 25% of the shares were sold to the public on the Nairobi Securities Exchange (NSE) in 2008 (Privatization Commission, n.d.). Vodafone Kenya, which was recently sold by the Government 15%, is now owned by Vodacom, which now holds 55%; the Government of Kenya holds 20%, while institutional and retail investors hold 25%. (People Daily, 2026) Safaricom is a mobile network provider that is the largest one in Kenya; it provides the services of voice, data, and mobile financial services through its leading product, M-Pesa (Kihiko et al., 2024). In the company's 2026 annual report, the company noted a total of 57.9 million 90-day active customers in Kenya, approximately 46.8 million 90-Day active customers with M-PESA, 333,011 M-PESA agents, and 5,648 employees (Safaricom PLC, 2026). In the same year, M-PESA reported transactions of KShs 41.68 trillion, while Safaricom PLC reported transactions valued at KShs 105.28 trillion and accounted for an estimated 65.7% of Kenya's telecommunications market (CA, 2024). The two repercussions of this scale on the company's reputation are that it is considered negative and devastating. Kenya is a unique risk environment because many people rely on mobile services,

ST. PAUL'S UNIVERSITY

some lack access to financial institutions, and the social media landscape is rapidly evolving, with limited institutional fact-checking (CA, 2024; Ndemo & Weiss, 2017).

In the second case, the company's reputation has become a commercial asset because it's now a financial utility, and Safaricom has become a commercial good as well as a regulatory and public responsibility (Safaricom Annual Report, 2024; FSD Kenya, 2023). Customer confidence, retention, and profitability are therefore directly impacted by corporate image (Murimi & Anyieni, 2024). The company's high visibility has also made it a common target of misinformation, such as fake promotional messages, fraudulent M-Pesa alerts, fake system outage notifications, and social media accounts claiming to be Safaricom. Safaricom has taken the bulk of the blame, responding on its own platform. It is also informing its customers about the fraud awareness page, which advises them that the number used by them for calls is 0722000000 only, while their PIN should not be shared with anyone, and they should forward the fake messages to 333 (Safaricom PLC, n.d.-a). Customers are also able to lodge complaints of suspected fraud using the mySafaricom app (Safaricom PLC, 2026, April). It advises consumers to check the official website and social media outlets, and warns about consumers reporting impostors, among other tips to avoid fake branded links (Safaricom PLC, n.d.). The guidance is based on the concepts of conventional fraud and impersonation. The issue this study is examining is whether these are sufficient countermeasures to deal with misinformation generated by AI. Unresolved incidents result in confusion to stakeholders, loss of brand loyalty, and a loss of reputation (Nuortimo et al., 2024; Murimi & Anyieni, 2024).

ST. PAUL'S UNIVERSITY

1.1 Statement of the problem

Artificial intelligence (AI) has fundamentally reshaped the corporate information landscape. It allows the rapid creation and mass dissemination of synthetic misinformation, such as deepfakes, manufactured press releases, and fake financial announcements, which organizations find hard to detect and combat in real time (Fung, 2025; Ma et al., 2025). The most important asset of corporations is reputation, which directly affects customer loyalty and investor confidence, and the organization's work is severely exposed in this climate (Nuortimo et al., 2024). Although ready-to-use models of crisis communication, such as the Situational Crisis Communication Theory (SCCT) by Coombs (2007) and the image repair theory of Benoit (1997) provide a set of guidelines on how reputational crises should be handled, they

were originally designed to address what are considered to be conventional crisis communication crises and have not been systematically evaluated against the problem of AI-driven misinformation, a gap (Upadhyay & Upadhyay, 2023).

This is a big challenge for Safaricom PLC in particular. It has established a strong foothold in the market through M-Pesa, a financial utility, and the Kenyan information environment moves so fast that it leaves little room for fact-checking (CA, 2024; Safaricom Annual Report, 2024; FSD Kenya, 2023). Safaricom PLC sits at its center. Commanding over 65.7% of the telecommunications market and operating M-Pesa, which processed over KES 36.8 trillion in transactions in 2023/2024, the company functions as a de facto financial utility whose reputational integrity is simultaneously a commercial asset, a regulatory obligation, and a public good (Safaricom Annual Report, 2024; FSD Kenya, 2023). AI-generated misinformation targeting Safaricom, therefore, carries the potential to destabilize critical national financial infrastructure, not merely damage a corporate brand. This risk is not hypothetical: in 2024, data privacy allegations materially damaged the company's public image amid heightened political tensions that amplified public receptivity to negative narratives (KHRC, 2024). Despite existing fraud awareness and crisis response programs, no empirical evidence establishes whether these interventions are adequate against AI-generated fake news (Murimi & Anyieni, 2024). This study addresses this dual theoretical and empirical gap by examining how Safaricom PLC deploys crisis communication strategies in response to AI-driven misinformation.

ST. PAUL'S UNIVERSITY

1.4 Purpose of the Study

The purpose of study is to examine the crisis communication strategies employed by Safaricom PLC in response to AI-generated misinformation and to assess their effectiveness in protecting and restoring corporate reputation.

1.5 Objectives

i. To identify the typologies and propagation patterns of AI-generated misinformation targeting Safaricom PLC.
ii. To examine the crisis communication strategies deployed by Safaricom PLC in response to AI-generated misinformation.

iii. To assess the effectiveness of Safaricom PLC's crisis communication responses in protecting and restoring corporate reputation in the context of AI-driven misinformation incidents.

iv. To determine the influence of AI-misinformation-triggered crisis communication on stakeholder trust.

1.6 Research questions

i. What are the predominant typologies and propagation patterns of AI-generated misinformation targeting Safaricom PLC?

ii. What crisis communication strategies has Safaricom PLC deployed in response to AI-generated misinformation?

iii. How effective are Safaricom PLC's crisis communication responses in protecting and restoring corporate reputation in the context of AI-generated misinformation incidents?

iv. How does Safaricom PLC's crisis communication response to AI-generated misinformation influence stakeholder trust?

1.7 Justification of the study

The emergence of AI and its widespread use are realities modern society cannot wish away. The current study is critical because it helps address theoretical gaps by examining existing crisis communication frameworks that were not developed to address AI-driven misinformation (Huang et al., 2026). Empirically, the study will help fill gaps in research on AI-driven misinformation, as the majority of the studies focus on crisis communication before the problem of AI misinformation (Arcuri et al., 2023). Additionally, Safaricom PLC was deemed appropriate for the current study as it is the largest tech corporation in the country that is vulnerable to AI misinformation. In practical terms, the study will be valuable for Safaricom PLC, as it will help the company adopt crisis communication strategies that can work effectively within the company and reduce the risks associated with artificial intelligence-mediated crises.

1.8 Significance of the study

The study contributes to research, policy, and practice. First, the current study will contribute to the emerging body of research on communication crises in the age of artificial

ST. PAUL'S UNIVERSITY

intelligence. In practice, the study may provide Safaricom and other organizations with evidence-based guidance to strengthen their reputation management strategies. For practitioners, the study provides a contextually grounded reference for AI-mediated crisis communication. Lastly, at the policy level, the study's findings can inform regulatory frameworks for digital misinformation in the Kenyan telecommunications sector.

1.9 Assumptions

Several assumptions were made in the study. First, it was assumed that the key informants, who were drawn from Safaricom PLC, external stakeholders, and expert informants, would respond honestly and give an accurate picture of their experience and perception of the use of AI to spread misinformation and what Safaricom PLC was doing about the crisis communication. Second, it was assumed that the documents and social media content analyzed were representative of the communications, incidents, and stakeholder reaction as they claimed to be. Third, the study found that Safaricom PLC, being Kenya's largest telecoms and mobile-money service provider, was a suitable and informative case to study the dynamics of AI-driven misinformation and crisis communication. Fourth, participants were assumed to have a fairly stable sense of what constitutes a piece of misinformation or an AI-generated piece of information, which is consistent with the operational definitions used in this study. Lastly, and perhaps most importantly, the study relied on the existence of two theories, the Situational Crisis Communication Theory and the Image Repair Theory, that were developed before the advent of generative AI, but were still seen as powerful analytical tools to probe crisis communication responses to reputational crises caused by AI.

1.10 Scope of the study

Conceptually, the study will be limited to AI-driven fake news, crisis communication, and corporate reputation. Geographically, the study will be limited to Kenya, focusing on Saffarciom PLC and the period from 2021 to 2026. This period saw significant advancement and escalation of misinformation risks as the use of AI technology became more mainstream

1.11 Limitations and Delimitations of the Study

The study has the potential to experience the following limitations. First, the single case study design might limit the generalizability of the findings of the study as it focuses on a single company (Taherdoost, 2022). Secondly, the study adopted a non-probabilistic

ST. PAUL'S UNIVERSITY

sampling technique, which might not produce a representative sample. Moreover, artificial intelligence has been evolving rapidly compared to company policies. Lastly, since the study is primarily qualitative and will involve the interviewers, there is likely to be a risk of social desirability bias that might influence participants' responses. The study was intentionally delimited to Safaricom PLC, selected for its dominant market position and the systemic significance of M-Pesa, making it the most contextually appropriate unit of analysis. Purposive sampling was deliberately adopted to prioritise depth of insight over statistical representativeness, consistent with qualitative inquiry. The study is temporally bounded to 2020–2024, capturing the mainstreaming of generative AI without claiming to account for developments beyond this window. Participants are limited to professionals with direct involvement in crisis communication, whose occupational grounding is expected to minimise social desirability bias and anchor responses in organisational practice.

1.12 Definition of terms

AI-generated misinformation: Misinformation content (including text, image, audio or video) created or amplified with the help of artificial intelligence tools, that appears to be real (Ma et al., 2025).

Crisis communication: It involves the strategic exchange of information between an organisation and its stakeholders during the pre-crisis, crisis and after-crisis phases to minimise harm and restore the confidence of stakeholders (Seeger, 2022; Upadhyay & Upadhyay, 2023).

Corporate reputation: The overall reputation the stakeholders are holding of the organisation, which is built through the communication and past behaviour of the organisation, and is related to its trust and perceived loyalty (Nuortimo et al., 2024).

Deepfake: A synthetic auditory, image or video product created using AI technologies that mimics a person saying or doing something that didn't actually happen (Vlaeminck, 2023).

Trust by stakeholders: Customer, investor and regulator trust in the reliability, honesty and competence of an organisation (Murimi & Anyieni, 2024).

Image repair: Organisational response through communicative channels, including denial, avoidance of responsibility, minimization and rectification, mortification.

1.13 Summary of the study

This chapter has set the ground for the study by providing the background to AI-driven misinformation as an emerging threat to corporate reputations, specifically for Safaricom PLC

ST. PAUL'S UNIVERSITY

in the mobile-first, financially interdependent economy of Kenya. It introduced the problem under consideration, showcasing the existing theoretical and empirical issues with using crisis communication frameworks that were created before the emergence of generative AI. Research questions, purpose, and objectives were detailed, and the research's theoretical justification and significance were explained. Additionally, the chapter specified the scope of this research, recognized the delimitation and limitations, explained the assumptions made for this study and explained some important key operating terms. The relevant theoretical and general literature, as well as the empirical literature that underlie the study, are now reviewed in Chapter Two.

ST. PAUL'S UNIVERSITY

CHAPTER TWO: LITERATURE REVIEW

2.0 Introduction

This chapter gives an extensive overview of the previously written material pertinent to the phenomenon of interest: how AI-generated misinformation and corporate reputation management (CRM) intersect in the field of crisis communication practice. The review is organized so that it progresses from general conceptual issues to the specific empirical and theoretical issues of the study. It begins by conducting a broad literature review that charts the broader research agenda related to misinformation, digital reputation, and organizational communication in the era of generative AI. This is followed by an empirical review of the literature, which is organized thematically as corporate reputation management in the digital era, crisis communication strategies, stakeholder trust and perception during crisis situations, the unique case of AI misinformation from an African perspective and specifically the Kenyan perspective, and Safaricom PLC's specific vulnerabilities to reputational threats in the Kenyan setting. The study then moves to a discussion of the theoretical underpinnings that provide the foundation for the study: Situational Crisis Communication Theory (SCCT) and Image Repair Theory (IRT) as the two dominant theories through which corporate crisis response has historically been examined. The conceptual framework that connects these theories to the variables examined is then explored. Finally, the chapter summarises the research gap found in the literature that was reviewed and briefly synthesizes the directions of the solution that are outlined in Chapter Three.

2.1 General literature review

The spread of false information has become a major issue for corporations and businesses in general. Misinformation can harm the organisation by eroding consumer trust and ultimately affecting its bottom line (Hodge, 2025). Until recently, the chances of companies suffering damages as a result of misinformation campaigns were low, until the emergence of artificial intelligence (AI). In the current age, businesses and companies face three problems. First, the internet allows falsehoods to spread quickly, while artificial intelligence can replicate them and make them easily believable (Bethany, 2025). Additionally, people do not have to believe the rumours or stories to hurt the corporate image and financial bottom line; they just need to react to them. Fleming (2024) observes that cheap generative AI tools have lowered the barriers to creating and spreading disinformation, both in terms of cost

ST. PAUL'S UNIVERSITY

and manpower, resulting in misleading content being churned out with little human intervention.

Studies and media practitioners have argued that AI remains the most significant threat of modern times. According to an interview conducted by the Colombia Business School, the emergence of AI has seen the amount of misinformation increase drastically, and the risks of these showing up on websites that can negatively affect the reputation of businesses have equally increased (Sperling, 2024). In 2024, misinformation was demonstrated in various incidents. For instance, there was a scam involving Elon Musk. One such deep fake involved Elon offering investment advice, which cost America billions of dollars (New et al., 2024). In the United Kingdom, Arup lost $25 million after an employee was tricked with a video call featuring a deepfake version of senior executives (Milmo, 2024). Some companies, including WPP, were lucky to detect the misinformation and thwart it before it became a major issue in the organization.

The advent of AI with its current capabilities has challenged the traditional crisis management system in the context of the media. New et al., (2024) observe that there is a need for organizations to be proactive in the detection and prevention of AI misinformation in the current digital age. Besides, there are calls for improved threat intelligence as the first step for organizations to prevent AI misinformation in the media setting (Kaul, 2025). One such approach is the human-in-the-loop approach. Unlike the media monitoring and social listening platforms in the pre-AI world, which relied on keyword matching, sentiment scoring, and volume-based alerts that worked to track brand and media mentions, there is a need to put a human in the loop who can validate what matters and remove all that does not matter while also prioritizing the risk responses.

In Africa, the impact of AI misinformation has largely been felt on the political front. Since the emergence of AI. The current data shows that AI misinformation has been critical in driving the political agenda (Musau, 2026). In Sudan, AI voice cloning has emerged as a significant threat to the peace of the region. Goodman & Hashim (2023) observe that there has been a campaign using artificial intelligence to impersonate Omar al-Bashir, the former leader of Sudan, and it has received hundreds of thousands of views on TikTok, something that has resulted in greater confusion in the country and further fueled the civil war that is currently

ST. PAUL'S UNIVERSITY

going on. The trend of misinformation is not just in Sudan; in Nigeria, the Minister of Public Information has decried the use of AI in misrepresenting public officials (Anyanwu, 2025). This move in itself undermines the public trust in the government, something that can result in adverse political outcomes and also shape the public opinion on various civic issues

In Kenya, the principal secretary for foreign affairs apologized for sharing a deepfake video on his social media platforms (Kimuyu, 2025). These trends highlight how much of a challenge AI misinformation has become in Africa, especially on the political front. Reports by Anthropic indicate that AI is being used to drive public sentiment on social media. The report indicates that the company foiled a plan by a social media actor to generate content promoting the cabinet secretary for energy amid the looming fuel and electricity crisis (Anthropic, 2026). Moreover, there are concerns that the use of AI to drive misinformation is likely to grow as the country enters an election period. Macharia, (2026) observes that the emergence of AI is likely to shape the 2027 elections. The author observes that the AI landscape in the country has largely been used to drive propaganda and shape narratives that can influence electoral outcomes. Nonetheless, despite the emergence of AI, the main focus has been on the political angle, with limited emphasis on the corporate setting; hence, the need to conduct research in the corporate context.

ST. PAUL'S UNIVERSITY

2.2 Empirical Literature Review

The proliferation of AI-generated misinformation represents one of the defining information challenges of the contemporary era. The rapid development of large language models such as GPT-4, Gemini, and Claude, alongside multimodal generative systems capable of producing synthetic images, audio, and video, has fundamentally democratised the capacity for misinformation production, placing highly sophisticated content-generation tools within reach of actors with minimal technical expertise. This democratisation has dramatically lowered the barriers to the creation of convincing fake content, enabling malicious actors to produce and disseminate false information at unprecedented speed and scale.

Scholars have identified several distinct mechanisms through which AI amplifies misinformation. First, AI systems enable the automated production of false content, allowing a single actor to generate thousands of unique but coordinated false narratives simultaneously (Romanishyn et al., 2025). Second, AI-powered micro-targeting allows disinformation to be

tailored to specific audience segments, increasing its emotional resonance and persuasive impact (Sonni, 2025). Third, algorithmic amplification on social media platforms driven by engagement-maximising recommendation systems systematically privileges emotionally arousing content, regardless of its veracity, thereby accelerating the diffusion of false narratives. Fourth, the visual verisimilitude of deepfake technology imbues fabricated content with an authority that text-based misinformation lacks, exploiting the well-documented human tendency to treat visual evidence as inherently more credible than verbal claims (Chesney & Citron, 2019).

The empirical evidence for the real-world consequences of AI-generated misinformation is substantial and growing. Vosoughi, Roy, and Aral's (2018) landmark study of Twitter, analyzing over 126,000 news stories shared over a decade, found that false information spread significantly faster, farther, and more broadly than accurate information across all categories. This finding is particularly consequential in the corporate context, where reputational damage can spread faster than organisations can respond. Ma et al. (2025) demonstrate that AI-generated fake financial announcements can cause measurable short-term disruptions in stock prices, while Arcuri et al. (2023) document that deepfakes and fabricated news stories have produced quantifiable volatility in both American and European Union equity markets.

From a cognitive psychology perspective, the illusory truth effect, whereby repeated exposure to a claim increases its perceived credibility regardless of factual accuracy, is particularly relevant to understanding why AI-driven misinformation poses such a persistent reputational challenge. In high-velocity disinformation environments, where false claims circulate repeatedly across multiple platforms, this effect can render initial corrections ineffective over time, as audiences are exposed to the false claim far more frequently than to corrective information. This phenomenon has profound implications for crisis communication strategy, suggesting that simple denial may be insufficient in the AI era and that organisations must engage in sustained, multi-platform corrective communication to counteract the illusory truth effect.

ST. PAUL'S UNIVERSITY

2.2.1 Corporate Reputation Management in the Digital Age

Corporate reputation has long been recognised as one of the most strategically significant intangible assets available to organisations. Fombrun (1996) defined reputation as the collective perception of a firm's past actions and prospects that describes the firm's overall appeal to all its key constituents when compared to its rivals. This conception has been widely adopted in the management literature, with subsequent scholars emphasising the multi-stakeholder nature of reputation and its direct influence on customer loyalty, investor confidence, employee retention, and regulatory goodwill (Nuortimo et al., 2024; Murimi & Anyieni, 2024).

The digital transformation of the media landscape has fundamentally altered the dynamics of corporate reputation management. Historically, organizations maintained substantial control over their public narratives through carefully managed relationships with mainstream media outlets and through the deliberate management of official communications channels. The emergence of social media platforms, user-generated content, and the 24-hour news cycle has eroded this control, creating an environment in which negative information can originate from virtually any source and spread globally within minutes (Coombs, 2014). The entry of AI into this landscape has further intensified this vulnerability, enabling sophisticated actors to produce and distribute highly credible false content that can rapidly contaminate an organization's information environment (Fung, 2025).

The telecommunications sector faces particularly acute reputational vulnerabilities in this context. Telecommunications companies occupy a unique position in the information ecosystem as the infrastructure providers through which digital communication flows, making them simultaneously central to misinformation diffusion and exposed to public scrutiny regarding their own conduct. Moreover, the financial services components of many telecommunications platforms, exemplified by Safaricom's M-Pesa, make them especially susceptible to misinformation relating to financial stability, data security, and service reliability, areas where stakeholder trust is both particularly important and particularly fragile (Murimi & Anyieni, 2024).

Research on the reputational consequences of crises consistently demonstrates that the duration and severity of reputational damage are shaped not only by the nature of the

ST. PAUL'S UNIVERSITY

precipitating event but also by the quality of the organizational response. Coombs and Holladay (2008) find that organizations that respond swiftly, transparently, and with apparent concern for affected stakeholders consistently suffer less severe and shorter-duration reputational damage than those that respond slowly, evasively, or defensively. This finding has important implications for the AI misinformation context, where the speed of false information diffusion may require organizations to respond with unprecedented rapidity, before the full scope of the misinformation incident is understood.

2.2.2 Crisis Communication Strategies: An Overview

The scholarly literature on crisis communication has produced a rich body of knowledge regarding the strategies organizations employ when their reputations are threatened. Beyond the SCCT and IRT frameworks already discussed, several additional bodies of work are relevant to this study. Frandsen and Johansen (2017) advance a rhetorical arena approach that conceptualizes crisis communication as occurring within a multi-vocal arena in which numerous actors, including organizations, media, public authorities, and affected publics, simultaneously engage in communicative acts that collectively shape crisis narratives. This perspective is particularly relevant to the AI disinformation context, where the organization is only one of many actors contributing to the informational environment.

The literature broadly distinguishes between reactive and proactive crisis communication approaches. Reactive approaches, which represent the dominant paradigm in both SCCT and IRT, focus on the strategies organizations deploy after a reputational threat has materialized. These include issuing timely public statements, engaging directly with affected stakeholders, correcting false information through authoritative channels, and, where appropriate, acknowledging shortcomings and committing to improvement (Seeger, 2022). Proactive approaches, by contrast, focus on the prevention or mitigation of reputational crises before they occur, through strategies such as continuous monitoring of the information environment, pre-emptive stakeholder communication, and the establishment of robust organizational credibility that can serve as a reputational buffer in times of crisis (Upadhyay & Upadhyay, 2023).

The growing significance of proactive crisis communication in the AI era is reflected in an emerging literature on "pre-bunking" and "inoculation theory" approaches to

ST. PAUL'S UNIVERSITY

misinformation management. Drawing on inoculation theory from social psychology, which posits that exposure to weakened forms of persuasive attacks can build resistance to more potent future attacks, scholars have proposed that organizations can inoculate their stakeholders against misinformation by proactively explaining the techniques commonly used in AI-generated fake content and encouraging critical media literacy. This approach represents a significant departure from traditional crisis communication frameworks, which are predominantly reactive in their orientation, and its applicability in the corporate telecommunications context warrants empirical investigation.

Digital communication channels have also substantially reshaped the landscape of crisis communication strategy (Nuortimo et al., 2024). Social media platforms, in particular, have become primary arenas for crisis communication, enabling organisations to respond rapidly and directly to stakeholder concerns without the mediating influence of mainstream journalism (Ennab et al., 2022) However, these platforms also present significant challenges for crisis communicators, as they are simultaneously the primary vectors through which AI-generated misinformation spreads and venues in which organisational communications must compete with user-generated content for audience attention. Research Xu (2020) demonstrates that the channel through which crisis information is communicated significantly influences stakeholder responses, with social media communications generating different patterns of information-sharing behaviour compared to more traditional channels, a finding with direct implications for the multi-channel communication strategies employed by organizations such as Safaricom PLC.

2.2.3 Stakeholder Trust and Perception in Crisis Contexts

Stakeholder trust is widely acknowledged in the crisis communication literature as both a central organizational resource and a primary casualty of reputational crises. Mayer, Davis, and Schoorman's (1995) influential model of organizational trust identifies ability, benevolence, and integrity as the three foundational dimensions of trust, with crises typically threatening one or more of these dimensions. In the context of AI-driven misinformation, ability and integrity are particularly vulnerable to attack, as fake news can fabricate evidence of organizational incompetence, financial instability, or ethical misconduct, potentially

ST. PAUL'S UNIVERSITY

undermining stakeholder confidence in the organization's core competencies and moral character simultaneously.

Empirical research consistently demonstrates that the quality of crisis communication has a significant direct effect on post-crisis stakeholder trust. Coombs and Holladay (2008) find that stakeholder goodwill accumulated through a history of positive organizational conduct and stakeholder-centered communication functions as a reputational "halo" that moderates the negative reputational impact of crises, particularly in cases where the organization is not genuinely responsible for the precipitating event. This finding has important implications for the current study: organizations with strong pre-crisis reputations, such as Safaricom PLC, which has cultivated significant brand equity over its two-decade history in the Kenyan market, may be relatively better positioned to weather AI-driven reputational attacks, provided their crisis communication responses are adequately managed.

Research on the role of transparency in crisis communication further supports the importance of open, honest communication in maintaining and restoring stakeholder trust. Rawlins (2008) argues that organizational transparency, defined as the proactive provision of accurate, complete, and timely information to stakeholders, is a foundational prerequisite for sustained trust, particularly in contexts where stakeholders have reason to doubt organizational communications. In the AI disinformation context, where the credibility of information itself is contested, organizational transparency takes on heightened significance: organizations that are perceived as consistently honest and forthcoming are more likely to be believed when they assert the falsity of AI-generated claims than those with a history of evasive or incomplete communication (Rawlins, 2008).

The intersecting roles of emotional communication and empathy in crisis response also emerge prominently from the literature. Hess and Huq (2014) find that crisis communications perceived by audiences as genuinely empathetic, demonstrating sincere concern for the welfare of affected parties, rather than merely managing organizational exposure, generate significantly more positive stakeholder evaluations than those perceived as primarily self-protective. This suggests that even in cases of AI-fabricated crises, where the organization bears no genuine responsibility, communications that demonstrate awareness of stakeholder

ST. PAUL'S UNIVERSITY

concern and commitment to stakeholder well-being are likely to be more effective than purely defensive communications focused on asserting organizational innocence.

2.2.4 AI Misinformation in the African and Kenyan Context

Scholarly attention to AI-driven misinformation in Africa, and in Sub-Saharan Africa in particular, has grown substantially in recent years. However, the corporate dimension of this phenomenon remains relatively understudied compared to its political and electoral manifestations. The broader context of AI misinformation in Africa is characterized by several distinctive features that differentiate it from the Northern contexts in which most foundational crisis communication theory was developed. These include relatively lower levels of formal media literacy, high mobile internet penetration combined with uneven access to authoritative information sources, a high reliance on informal social networks for news consumption, and significant linguistic and cultural diversity that complicates both the production and verification of content (Okolo, 2024).

In the Kenyan context specifically, AI-driven misinformation has manifested primarily in the political and electoral sphere, with documented incidents during the 2022 general elections including fabricated results, bot-driven coordinated inauthentic behaviour, and AI-generated deepfake audio and video of senior political figures (Miyandazi & Thuo, 2025; Kamau & Shiundu, 2024). These incidents provide important contextual insights for the corporate sector, as they demonstrate the technical sophistication of disinformation actors operating in Kenya, the vulnerability of Kenyan digital audiences to AI-generated content, and the limitations of current institutional responses to disinformation at scale.

Kenya's digital infrastructure presents a dual-edged reality for corporate communication. On one hand, the country's high mobile penetration rate, with smartphone penetration among the highest in Sub-Saharan Africa, and the widespread use of social media platforms, including Facebook, WhatsApp, X (formerly Twitter), and Instagram, provide organizations with multiple high-reach channels for crisis communication (Mwakideu, 2022). On the other hand, these same platforms serve as primary vectors for the rapid dissemination of misinformation, with Kamau and Shiundu (2024) documenting the extensive role of WhatsApp in particular in the spread of false information in the Kenyan context. The closed, encrypted nature of WhatsApp communications presents particular challenges for

ST. PAUL'S UNIVERSITY

organizational counter-messaging, as false information can circulate within private group networks that are inaccessible to organizational monitoring or correction.

The regulatory environment for digital misinformation in Kenya is still developing. While the Kenya Information and Communications Act and the Computer Misuse and Cybercrimes Act (2018) provide some legal basis for addressing online misinformation, enforcement remains uneven, and the technical sophistication of AI-generated content presents significant challenges for regulatory actors and fact-checking organizations operating with limited resources (Miyandazi & Thuo, 2025). This regulatory gap places additional communicative responsibility on organizations such as Safaricom PLC, which must navigate reputational crises in an environment characterized by both institutional vulnerability to AI misinformation and limited external support for corrective action.

2.2.5 Safaricom PLC: Reputation and Misinformation Exposure

Safaricom PLC occupies a distinctive position in the Kenyan corporate landscape, making it an especially instructive case for examining the intersection of AI-driven misinformation and corporate crisis communication. As Kenya's largest telecommunications company by subscriber base and market capitalization, and as the operator of M-Pesa, one of the world's most widely used mobile financial services platforms, Safaricom exercises substantial influence over Kenya's digital economy and occupies a position of significant public trust (Kihiko, Yatich & Obuba, 2024). This prominence confers reputational capital while also making the company a high-value target for reputational attacks.

Documented incidents of AI-assisted and digitally amplified misinformation targeting Safaricom include fraudulent M-Pesa promotional messages, false system outage announcements, fabricated customer service communications, and social media impersonation accounts that purport to represent the company (Murimi & Anyieni, 2024). These incidents align with the broader typology of corporate misinformation identified in the global literature, spanning false product or service announcements, fabricated financial communications, and impersonation of company representatives. The M-Pesa dimension is particularly significant: given the platform's role as a primary financial tool for millions of Kenyan households, misinformation about M-Pesa availability, transaction security, or user data privacy can cause both direct financial harm to consumers and severe reputational damage to Safaricom.

ST. PAUL'S UNIVERSITY

The 2024 data privacy controversy, in which Safaricom was accused of sharing customer data with government authorities in the politically charged context of anti-government protests, represents a particularly salient case for this study. The Kenya Human Rights Commission (KHRC, 2024) documented allegations that Safaricom's network infrastructure was used to facilitate surveillance of protest participants, allegations that the company disputed. This incident illustrates the complex intersection of genuine organizational controversy and AI-amplified misinformation in the Kenyan corporate context, as false and misleading claims circulated alongside legitimate reporting and civil society concerns, creating a highly complex informational environment for crisis communication management. Existing research on Safaricom's reputation management is largely focused on conventional brand management and customer satisfaction dimensions, with limited systematic attention to AI-specific misinformation and crisis communication (Murimi & Anyieni, 2024). This gap is significant given the company's prominence and the documented vulnerability of Kenya's digital information environment to AI-driven misinformation. The current study addresses this gap by providing the first systematic empirical examination of Safaricom's crisis communication strategies in the specific context of AI-generated fake news and their effects on stakeholder trust and corporate reputation.

ST. PAUL'S UNIVERSITY

2.3 Theoretical framework

Two theories will guide the study: Situation Crisis Communication Theory (SCCT) and Image Repair Theory.

2.3.1 Situation Crisis Communication Theory (SCCT)

Situational Crisis Communication Theory (SCCT), developed by Coombs (1995, 2007), is one of the most prominent theories in crisis communication research. Based on Weiner's (1985) work on Attribution Theory, SCCT assumes that when there is a crisis, stakeholders' attributions of responsibility directly influence their emotional and behavioural reactions to the crisis and the organizations involved. Numerous scholars have used SCCT as a predictive, evidence-based model to link the type of crisis to appropriate communicative responses, making it useful for analysing the use of AI-generated fake news as an organizational reputational threat.

The central tenet of SCCT is the notion of crisis responsibility, which is based on three situational factors: initial crisis responsibility, crisis history, and prior relational reputation. All three variables can be manipulated by AI-driven disinformation, with stakeholders' perceptions of the organization's responsibility, for instance, heightened, and the crisis being portrayed as a victim-cluster crisis to the public eye, rather than a preventable-cluster crisis. While not a form of actual misdeed, coordinated disinformation campaigns can also erode reputational capital, tainting the situational variables on which the prescriptive logic of SCCT is based.

There is a continuum of crisis response strategies, from denial to diminishment to rebuilding, as outlined by Coombs (2007) through the process of situational assessment. Rumours placed within the victim cluster, giving victims advice on how to deny false accusations. That prescription is, however, complicated by the "illusory truth effect described by Hasher et al. (1977) that exposure to a claim makes it seem more credible to those to whom it is presented, even in the absence of a factual foundation. An elephant can easily mistake its nose for its foot, or a lion for a mouse. In the high-velocity world of information, repeated denials can paradoxically exacerbate false information among audiences, creating a scenario where they see denials as concealment, a condition that is not well captured by the original SCCT formulation.

2.3.2 Image Repair Theory (IRT)

The second theory guiding this study is Benoit's (2000, 2008, 2013) Image Repair Theory (IRT), also known as Image Restoration Theory. Originally rooted in the rhetorical tradition, IRT was developed as a comprehensive typology of strategies that individuals and organizations deploy when their image is threatened or damaged. The theory is founded on two core assumptions: that communication is fundamentally goal-directed and that maintaining a positive reputation is a central organizational objective. Accordingly, when an actor's image is imperilled by an accusation of wrongdoing, whether real or perceived, they are compelled to engage in communicative acts designed to repair or restore that image (Benoit, 1997)

Benoit (1997) identifies five broad categories of image repair strategies, each containing several specific sub-strategies. The first category is denial, which encompasses simple denial, asserting that the act in question did not occur and shifting the blame to another party. The second category is evading responsibility, which includes provocation, defeasibility, accident, and good intentions. The third category is reducing the offensiveness of the event,

ST. PAUL'S UNIVERSITY

which includes bolstering, minimization, differentiation, transcendence, attacking the accuser, and compensation. The fourth category is corrective action, whereby the organization promises to rectify the problem. The fifth and final category is mortification, which involves a sincere apology and acknowledgement of wrongdoing (Benoit, 1997).

The relevance of IRT to the current study lies in its explicit focus on perception rather than factual truth. Benoit (1997) argues that an organization need not have actually committed the alleged wrongdoing for image damage to occur; it is sufficient that the relevant audience believes it to have done so. This perceptual basis renders IRT particularly applicable to AI-driven disinformation scenarios, in which fabricated content can generate audience beliefs that are entirely disconnected from organizational conduct. In such contexts, the conventional IRT distinction between denial (applicable when the act did not occur) and mortification (applicable when the organization bears genuine responsibility) becomes conceptually strained, because the organization must simultaneously assert its innocence and engage in communication that meaningfully addresses stakeholder concerns without inadvertently legitimizing false claims.

Moreover, the speed and scale of AI-generated misinformation challenge IRT's implicit assumption that organizations have adequate time to formulate and disseminate considered responses. In the current digital information environment, characterized by algorithmic amplification and viral information diffusion, reputational damage can accumulate within hours, outpacing the deliberate rhetorical processes that Benoit's framework presupposes. This temporal tension represents a significant boundary condition of IRT in the AI disinformation era, which this study seeks to explore in the context of Safaricom PLC. Both SCCT and IRT thus serve not as immutable prescriptions but as analytical frameworks whose assumptions and prescriptions are critically interrogated in light of the emerging dynamics of AI-mediated reputational crises.

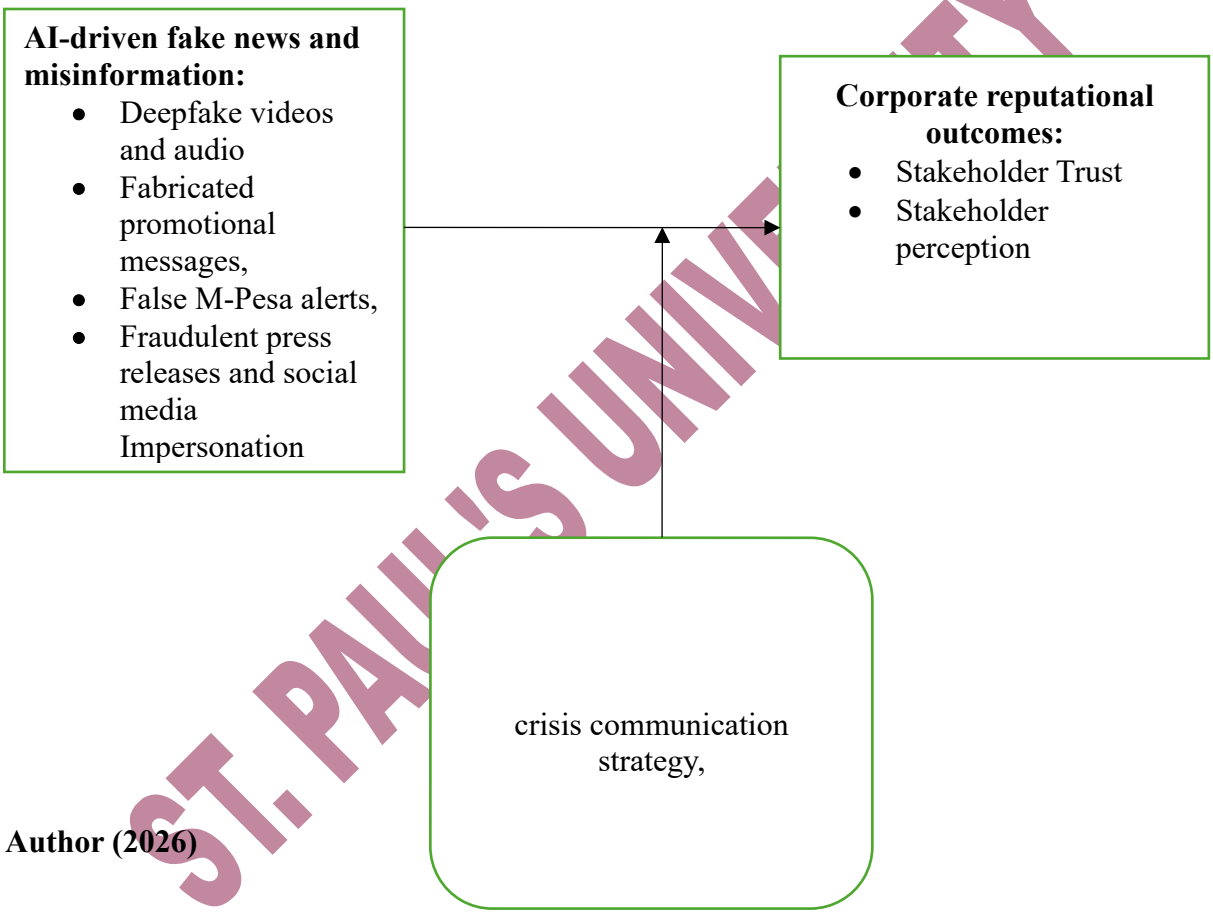
ST. PAUL'S UNIVERSITY

2.5 Conceptual Framework

The study investigated the role of the volume, velocity, credibility, and motivational source of AI-generated fake news and misinformation, like deepfakes, fake SMS news notifications, fake press releases, and spoofed accounts, and how these factors influence the crisis communication challenge. The crisis communication strategy is conceptualised as three

dimensions, namely: strategy type, channel, and timing, based on SCCT and IRT models. Safaricom's history tempers this effectiveness, the digital information context in Kenya, the media literacy of stakeholders, and the regulatory context. The dependent variables included perceived trustworthiness, competence, and integrity of corporate reputational outcomes and self-reported evaluations of Safaricom's responses and their relational consequences. These together embody reputation as an organizational asset and a relationship-based process.

Figure 2.1: Conceptual Framework



Author (2026)

ST. PAUL'S UNIVERSITY

2.4 Summary of Research Gap

First, the existing crisis communication literature suffers from a theoretical gap based on the underlying assumptions of the two major paradigms in this field. These theories, SCCT (Coombs, 2007) and image repair theory (Benoit, 1997), were created in a traditional media context that features visible human actors, a known source of crisis information, and a more

linear timeline between the emergence of a crisis and the response of the organization. Their core prescriptions, such as responsibility attribution and denial strategies, bolstering, mortification, etc., presume certain conditions that are structurally undermined by AI-generated misinformation. Deepfakes, fake press releases, and algorithmically churned financial releases add an extra dimension to the crisis because the authorship of the content is intentionally disguised, the content is indistinguishable from the original, and there is a speed dimension that surpasses the capacity of institutions to respond (Fung, 2025; Ma et al., 2025). However, the basic premise of both models, the ability of the organisation to detect a trigger or set of triggers for a crisis, to take responsibility for managing it and respond within a meaningful time frame, is not guaranteed in misinformation environments powered by Artificial Intelligence (Upadhyay & Upadhyay, 2023). While emerging scholarship has started to challenge SCCT in the digital and platform-mediated contexts (Austin & Jin, 2018; Coombs & Holladay, 2023), and some have specifically marked the absence of existing frameworks in dealing with synthetic media threats (Vaccari & Chadwick, 2020), there is no study yet that has methodologically examined if and how SCCT, or image repair theory, still has prescriptive value in situations involving AI-generated misinformation undermining corporate reputation. This study directly addresses this gap by examining both theories through an empirical case involving a reputational crisis managed with AI tools, to develop theories of crisis communication in the synthetic media age.

Second, there is a gap in empirical literature on corporate crisis communication in the context of AI-enabled misinformation in Sub-Saharan Africa, and specifically in the country of Kenya. There has been a significant expansion of the empirical evidence about the consequences of misinformation in the global literature, including a body of work on its reputational, financial and political impact in North American and European contexts (Chesney & Citron, 2019; Hancock et al., 2021; Weidman & Shell, 2021). There are, however, fundamental structural differences in Africa that create unique forms of misinformation risks and communicative tools for addressing them, such as the high uptake of mobile internet services, low institutional capabilities in fact-checking, an underbanked population for whom mobile money is a core financial service, and fast-paced social media diffusion (Ndemo & Weiss, 2017; CA, 2024). The limited empirical research on misinformation in Kenya focuses on political and electoral aspects (Miyandazi & Thuo, 2025; Kamau & Shiundu, 2024), with

ST. PAUL'S UNIVERSITY

little or no research on the communication dimension of misinformation in the corporate communication sector. Research on reputational crisis in organizations in Kenya has mainly centred on the traditional typologies of crisis rather than threats emanating from AI, and there has been limited research to gauge if current interventions used in the corporate crisis communication framework are sufficient in the context of synthetic misinformation. This is not just an academic issue, but one that Kenya is already facing today. Kenyan organisations, such as Safaricom PLC, for example, have had to deal with AI-related reputation crises without contextually appropriate, evidence-based guidance, specifically in 2024, with allegations of data privacy abuse that materially harmed Safaricom's reputation during a politically charged period of misinformation (KHRC, 2024). The current study attempts to fill this void by drawing contextually specific and empirically valid information from the Kenyan corporate communication context.

Third, the gap is a sector-specific gap involving the telecommunications sector, which has a very special place in the literature of crisis communication, and to date remains somewhat under-theorised. The role of telecommunications companies is a tripartite model; as such, they are also providers of digital infrastructure, platforms for financial services, and as primary channels of misinformation dissemination, which means they present communication challenges unlike other organisations in other industry sectors (Nuortimo et al., 2024). The audience most vulnerable to receiving misinformation on a telecom company are the ones that are also the company's customers, its most important stakeholders, and the chief agents of spreading misinformation further. Existing crisis communication frameworks generally do not consider that audiences play an active role in the propagation of a crisis; the frameworks focus on how one's reputation can be damaged by their own organisation (Coombs, 2019; Jin et al., 2014). At Safaricom PLC, this relationship has other systemic ramifications beyond the brand; an issue of reputational erosion equates to loss of trust in the M-Pesa service, which is the backbone of the service's effectiveness; without trust, users are unlikely to use the service. (FSD Kenya, 2023; Safaricom Annual Report, 2024). In the absence of sector-specific crisis communication frameworks for telecommunications companies threatened by misinformation stemming from the influence of AI, this study takes a step forward by providing insights from an empirical perspective that reflects the actual operations of a big telecom and mobile money company.

ST. PAUL'S UNIVERSITY

Fourth, the current empirical AI misinformation literature on corporate reputation is predominantly quantitative and experimental, which is a methodological gap. Fourth, the existing empirical literature on AI misinformation and corporate reputation is highly quantitative and experimental, a methodological gap. Event-study methodologies (Weidman & Shell, 2021), experimental exposure designs (Hancock et al., 2021), and large-scale content analyses (Vaccari & Chadwick, 2020) have yielded valuable evidence surrounding the aggregate reputational effects of misinformation and the measurable effects of various response strategies. However, these strategies do not reflect how and why responses to crisis communication can be effective in practice, rather than ineffective (Daymon & Holloway, 2011; Lindlof & Taylor, 2019). The experiential and processual nature of the process, the way communication professionals interpret and react to an event involving AI-driven misinformation, the influence of organisational hierarchy on decision-making around a crisis, how uncertainty is dealt with during a crisis, and how the effectiveness of a communication is assessed internally, is underrepresented in the quantitative literature. This methodological shortcoming is even more significant when considering the role of communication in organisations in Africa, which differ considerably from those in the North, where most quantitative research has been carried out, in terms of culture, institutional trust patterns, and platform affordances (Mutsvairo, 2016). To effectively fill this gap, this study employs a qualitative research methodology, involving in-depth interviews with communication practitioners and documentary analysis, to provide rich contextual insights of crisis communication experience in an AI era beyond the quantitative evidence that is available.

2.6 Summary of the chapter

The chapter provides a detailed review of the existing literature on AI and the Communication Crisis. The chapter provides a detailed exploration of the Situation Crisis Communication Theory (SCCT) and Image Repair Theory, and the research gaps that have been identified in the study.

ST. PAUL'S UNIVERSITY

CHAPTER THREE: METHODOLOGY

3.0 Introduction

This chapter outlines the methodological approach used to explore Safaricom PLC's crisis communication strategies in managing misinformation related to AI technology and explains each choice in terms of the study's interpretivist approach and exploratory goals. The chapter begins by explaining the research philosophy for the inquiry, then briefly outlines the research method and design, which in this case is an interpretivist single instrumental case study. It then explains the target population from which participants were selected, and the sampling method employed in identifying and selecting key informants from amongst Safaricom staff, external stakeholders, and independent experts. The chapter then presents details of the data collection methods employed, including key informants, documentary analysis, and analysis of social media content, and explains the data analysis approach adopted – the 6-step thematic approach developed by Braun and Clarke. The chapter also covers how trustworthiness and rigour were promoted in the development of the research; ethics issues involved in the conduct of the research; and concludes with a chapter summary that sets the stage for the presentation of the findings in Chapter Four.

3.1 Research Philosophy

Research philosophy refers to a system of beliefs and assumptions about how knowledge is developed and how researchers understand, design, and conduct their research (Bryman, 2012). Two primary components guide the research philosophy. First, the ontology, which studies the nature of reality, and epistemology, which concerns what we know and how we know it (May, 2025). The study was guided by the interpretivist research philosophy that holds that reality is not objective, but rather constructed through the meaning that individuals and groups attach to their experiences and actions (Saunders et al., 2009). This philosophical position suits the current study because it examines the effectiveness of crisis communication strategies in the context of AI-driven misinformation, which is inherently subjective and context-based. Corporate reputation, stakeholders' trust, and communicative efficiency are not measurable purely on objective terms but through interpretive processes of the stakeholders involved, whose perceptions are shaped by cultural, social, and organizational contexts that must be understood from within rather than observed from the outside.

ST. PAUL'S UNIVERSITY

Besides, the philosophy is appropriate considering the study is exploratory in nature. The limited research with regard to the intersection of AI misinformation and corporate crisis highlights that the phenomenon has not been systematically studied in the Kenyan telecom context, indicating that causal and conceptual categories at play are yet to be sufficiently understood, allowing for purely inductive inquiry. An interpretivist approach that is founded on inductive meaning-making and contextual understanding is hence epistemologically appropriate. Moreover, the study adopted a constructivist ontology by acknowledging that the phenomena of interest, such as corporate reputation, stakeholder trust, and communicative effectiveness, are socially constructed realities that are a product of organizational context, history, and cultural contexts in which the organization is embedded (Zubkov & Pondi, 2022). The study will hence be guided by interpretivist epistemology and constructivist ontology.

3.2 Research Methodology

Moreover, the study adopted a qualitative research methodology as the study sought to investigate a contemporary phenomenon in depth within its real-life context, as the study puts emphasis on Safaricom PLC's crisis communication response to AI-driven misinformation and fake news that are context-embedded phenomena, making the qualitative approach the natural methodological choice. Besides, the study sought relativity in meaning, multiplicity of interpretation, particularity, detail, and flexibility, which is at the core of the qualitative approach (Jwan & Ong'ondo, 2011). The study investigated AI fake news and misinformation in the Kenyan corporate setting. A qualitative approach enabled the study to ask questions that cannot be measured in figures through quantitative techniques. This perspective is consistent with the position of Alejandro and Zhao (2023), who indicated that qualitative research focuses on meaning-making and understanding a phenomenon. Moreover, the approach allowed for the triangulation of data collection strategies by combining key informant interviews, documentary analysis, and social media content analysis. According to Denzin (2012), triangulation involves the use of multiple data sources to cross-validate findings. It is a well-established approach for enhancing the trustworthiness of qualitative case study research. In the present study, triangulation is key, considering the phenomenon under investigation is multidimensional, encompassing organizational communication practices and stakeholder perceptions.

ST. PAUL'S UNIVERSITY

3.3 Research Design

The study adopted a single instrumental case study design. The instrumental case study design allowed the study to examine one sample and case, which allowed the study to develop a perspective related to the phenomenon being investigated. In the current study, the study selected Safaricom PLC, a telecommunications corporation, was selected to help in understanding the trends of misinformation in the Kenyan corporate setting and the telecommunications industry in general. In this context, Safaricom was selected as a case study to understand AI misinformation and corporate reputation. The choice of Safaricom is primarily due to its prominence not only in Kenya but also across borders, with M-Pesa operations in Ethiopia and other African countries. As Kenya's largest telecommunications company, it has a wide reach across diverse demographics, making it ideal for understanding AI and misinformation. Additionally, Safaricom was also considered appropriate for the study as it can be considered a "critical case" in the sense that it is uniquely positioned to illuminate the phenomenon of interest by virtue of market dominance, public visibility, and documented exposure to AI-driven misinformation (Diem et al., 2023). Secondly, the depth of analysis required addresses all four research objectives better served by a concentrated single case design rather than a comparative multi-case approach that would sacrifice analytical depth for breadth. Lastly, the current resource constraints favour a single-case approach that allows comprehensive data collection and analysis within the available parameters.

3.4 Population

The population of a study is the total number of people who have the characteristics of interest to the study (Battista & Torre, 2023). This study had three populations. The first one was all Safaricom PLC employees in Kenya. This workforce includes members of the corporate, communications, digital media, and brand management teams that make up the 6800 employees at Safaricom PLC. The second was the external stakeholders. It comprised 57.9 million 90-day active Safaricom customers, of whom approximately 46.8 million were M-PESA users (Safaricom PLC, 2026). It also included 12,162 registered NGOs (National Register of NGOs, Media Council of Kenya, 2025) and journalists covering business and technology, all registered by the Media Council of Kenya. The third was professionals in AI, misinformation, crisis communication, and reputation management in Kenya. These included

ST. PAUL'S UNIVERSITY

academics and experts on digital and AI misinformation. At the Communications Authority of kenya the population is estimated to be 120 employees.

3.5 Target Population

The target population of a survey denotes the specific units from which the survey data are to be used to make inferences (Battista & Torre, 2023). The first group comprises Safaricom PLC communication and corporate affairs professionals (N=45), including individuals directly responsible for crisis communication, public relations, social media management, and brand reputation. This group is uniquely positioned to provide authoritative and in-depth accounts of the company's crisis communication strategies, decision-making processes, and institutional responses to AI-driven misinformation incidents. The second group consists of external stakeholders, including active Safaricom customers and civil society actors who have directly encountered or engaged with Safaricom's crisis communication efforts (N=.20). This group provides an external stakeholder perspective essential for assessing how Safaricom's communication strategies are received and evaluated by those they are designed to reach and reassure. The third group includes expert informants, comprising academics, digital communication specialists, and regulatory professionals with recognized expertise (N=30) in AI misinformation, crisis communication, or corporate reputation management in Kenya. This group contributes analytical depth and contextual insight that complements the organizational and stakeholder perspectives.

ST. PAUL'S UNIVERSITY

3.6 Sampling Technique

This study employed purposive sampling as its primary sampling strategy. Purposive sampling involves the deliberate selection of participants based on specific, predefined criteria relevant to the research objectives to maximise the information richness of the data collected rather than to achieve statistical representativeness (Ames et al., 2019). This approach is standard and appropriate in qualitative case study research, where the objective is analytical generalisation rather than statistical generalisation, and where the depth and relevance of participant knowledge are more important than sample size or random selection.

For Safaricom communication professionals, the selection criteria included: current or recent employment in a communication, public relations, corporate affairs, digital media, or brand management role at Safaricom PLC; a minimum of two years of relevant experience;

and direct or indirect involvement in the organisation's response to misinformation or reputational incidents. Approximately eight to ten participants were recruited from this group. For external stakeholders, the criteria included direct experience of Safaricom-related AI misinformation incidents, active engagement with Safaricom's communication channels, and willingness to provide candid assessments of the company's crisis communication responses; ten to twelve participants were recruited. For expert informants, the criteria included recognised expertise in digital communication, AI misinformation, or corporate reputation management in Kenya, and a minimum of three years of relevant professional or academic experience; five to seven expert informants will be recruited.

Snowball sampling was employed as a supplementary strategy within each participant group, with initial participants invited to refer the study to other individuals meeting the selection criteria (Bryman, 2016). Data collection continued until theoretical saturation, defined as the point at which new data no longer generate new conceptual insights, is reached (Glaser & Strauss, 1967). Based on comparable qualitative studies in the crisis communication literature, saturation was anticipated within a total participant pool of approximately twenty-five to thirty individuals across all three groups.

ST. PAUL'S UNIVERSITY

3.7 Sample size

The sample size was determined using data and theoretical saturation. While most qualitative studies rely on data saturation, according to Guest et al. (2006) a study involving 60 interviews, the elements and saturation emerged within the first six. The assertion is consistent with Hennink et al., (2017) the observation of data saturation in the ninth interview in a study involving 25 participants. The rapid plateau creates an illusion that the sample is adequate even when deeper theoretical exploration is not achieved. To overcome this challenge, the use of theoretical saturation, which relates to the point at which further data collection fails to add to the explanatory power of the emerging theory (Gandy, 2024). Integrating data and theoretical saturation enabled the study to shift from the mere absence of new information to the theoretical value of additional information. The current study hence targets between 20 and 25-30 participants from the different categories

Table 3.1 Population and Sample size

St. Paul's University Repository

Participant group	Target population description	Sample size
Group 1: Safaricom Communication & Corporate Affairs Professionals	Corporate Affairs & Communications division staff	**8–10**
	Digital Communications & Social Media team members	Incl. above
	Customer Experience & Brand Management professionals	Incl. above
Group 1 Sub-total	Corporate Affairs, Digital Communications & Brand Management staff	**8–10**
Group 2 External Stakeholders	Active Safaricom subscribers (Kenya) who have encountered misinformation about Safaricom/M-Pesa	**7–8**
	Civil society representatives (NGOs, consumer rights bodies) engaged with Safaricom's corporate conduct	**2–3**
	Media professionals (journalists, editors) who have reported on Safaricom misinformation incidents	**1–2**
Group 2 Sub-total	Safaricom customers, civil society actors & media professionals	**10–12**
Group 3 Expert Informants	Digital communication & crisis PR academics/studys	**2–3**
	AI misinformation & digital media specialists	**2–3**
	Regulatory professionals	**1–2**
Group 3 Sub-total	Academics, digital communication specialists & regulatory professionals	**5–7**
TOTAL	**All three participant groups**	**25–30**

Author (2026)

ST. PAUL'S UNIVERSITY

3.8 Data Collection

The study adopted three complementary data collection tools. Key Informant interviews, documentary analysis, and social media content analysis. These tools enabled cross-validation of findings and the construction of a comprehensive, multi-perspective account of the phenomenon under investigation.

3.8.1 Key informant interviews

The study collected data from Safaricom communication professionals using the key informant interviews. These officers within the department of communication were purposively selected because they possess valuable, resource-rich information regarding the phenomenon under investigation. The method allowed the study to build rapport with the participants, enabling them to provide valuable and valid information regarding the study (Akhter, 2022). The interviews were open-ended to permit flexibility in the process. Interview guides will be the instrument used to collect the data.

3.8.3 Document Analysis

Document analysis constitutes the second data collection method. Documents provide a stable, independently verifiable record of organisational communication practices that complements and cross-validates interview participants' accounts, helping mitigate the risk of social desirability bias in self-reported interview data (Júnior, 2022). The documentary corpus for this study will encompass both internal organisational documents and publicly available materials. Internal documents will be sought from Safaricom PLC and may include: official crisis communication policies and guidelines; internal communication plans and response protocols for misinformation incidents; board and management communications relating to specific AI misinformation incidents; and training materials for staff on misinformation detection and response.

Publicly available documents included: Safaricom's official press releases and public statements relating to misinformation incidents during the 2021–2026 study period; Safaricom's annual and sustainability reports from 2023; media coverage of Safaricom misinformation incidents in Kenyan print, broadcast, and digital media; reports and analyses published by Kenyan civil society organisations, regulatory bodies, and fact-checking organisations relating to digital misinformation; and relevant policy documents from the Communications Authority of Kenya and other regulatory bodies. Documentary data was

ST. PAUL'S UNIVERSITY

collected through Safaricom's official website and media archive, the Communications Authority of Kenya's document repository, the Kenya Gazette, and established Kenyan media archives, including the Daily Nation, Business Daily, and The Standard digital archives, all within the period of 2023 to date.

Documents were assessed for authenticity, credibility, representativeness, and meaning using the evaluative criteria established by Scott (1990). Given the sensitivity of internal Safaricom documents, the study acknowledges that access may be limited, as noted in the study's delimitations. In such cases, publicly available materials and interview data will form the primary evidentiary basis for analysis, with internal documents treated as supplementary where available. All documentary data will be catalogued in a structured reference management system using Zotero, which facilitates systematic organisation, annotation, and retrieval of source materials throughout the analysis phase.

3.8.4 Social Media Content Analysis

The third method of data collection is social media content analysis. Considering that the social media platforms were the key source of AI-driven misinformation about Safaricom and the key outlets of Safaricom responding to the crisis situation, it is important to systematically explore the content of social media in order to capture the dynamics of crisis communication in the online setting. The approach provided empirical data on the nature and volume of AI-related misinformation circulating about Safaricom, Safaricom's official social media responses, the content of those responses, and how they were received, as indicated by audience attention and commentary.

The data was collected from these primary platforms: X (previously Twitter), Facebook, Instagram, and YouTube, which are the main official social media platforms of Safaricom, as well as the main platforms where misinformation about the company spreads in Kenya (Kamau & Shiundu, 2024). The scope of data collection was limited by the research period (2023 -2026). They included specific posts, threads, and comment sections regarding identified instances of AI misinformation involving Safaricom within the same timeframe. Brandwatch, a professional social media analysis and monitoring platform, was used to gather data because it allows keyword searches across multiple platforms, retrieves historical data, analyses sentiment, and tracks volume. The Boolean search features of Brandwatch will enable

ST. PAUL'S UNIVERSITY

the identification of specific posts with predetermined keywords related to Safaricom, M-Pesa, specific misinformation cases, and crisis communication events identified during the documentary analysis and interview stages. In the absence of Brandwatch data for specific platforms or timeframes, Crowdtangle (Facebook and Instagram) and hand-based searches on X using its advanced search features will be used as alternative data sources.

3.10 Trustworthiness and Rigour

The trustworthiness of qualitative research is typically evaluated against four criteria advanced by Lincoln and Guba (1985): credibility, transferability, dependability, and confirmability. This study addresses each criterion through specific methodological strategies. Credibility, the qualitative analogue of internal validity, is enhanced through member checking, whereby a summary of preliminary findings will be shared with a subset of interview participants for verification and feedback; through triangulation of data sources and methods as described above; and through prolonged engagement with the research site and participants throughout the data collection phase. Transferability, the qualitative analogue of external validity, is addressed through the provision of thick, rich descriptions of the research context, case, and findings, enabling readers to assess the applicability of the study's insights to other contexts (Geertz, 1973).

Dependability, the qualitative analogue of reliability, is ensured through the maintenance of a comprehensive audit trail documenting all methodological decisions, data collection procedures, analytical steps, and interpretive choices made throughout the research process (Bryman, 2012). This audit trail will be maintained in the study's reflexive journal in Microsoft OneNote and will be available for external review. Confirmability, the qualitative analogue of objectivity, is addressed through reflexive practice, whereby the study explicitly acknowledges and reflects on the potential influence of its own background, assumptions, and position on the research process and findings, and takes deliberate steps to ensure that findings are grounded in the data rather than in the study's preconceptions. Inter-coder reliability will be assessed for the thematic coding of interview transcripts, with a second coder independently coding a 20% sample of transcripts using the established coding framework; discrepancies will be resolved through discussion and framework refinement until acceptable agreement is reached.

ST. PAUL'S UNIVERSITY

3.11 Data Analysis

The data collected from the three methods were subjected to thematic analysis. This method involves identifying, analyzing, and reporting patterns of meaning within qualitative data. The study adopted this method for three main reasons. First, it is philosophically compatible with the study's intepretivist paradigm. Secondly, it is flexible enough to accommodate data generated from different data sources. Lastly, the design aligns with the study's objective to produce a theoretically informed interpretation that contributes to scholarship on crisis communication in the AI era. Thematic analysis supports the development of themes at both descriptive and interpretive levels of abstraction. The thematic analysis, as advanced by Braun and Clarke (2006), involves six steps. The first step includes familiarization with the data, which involves transcription and going through the information and data. The second step involves the generation of initial codes, which is followed by a search for themes, reviewing them, defining and naming them, and lastly generating the report. The findings are presented in narratives with verbatim quotes from the transcripts included in the findings of the study.

3.11 Ethical Considerations

The study sought permission from St Paul's University Ethical Review Board, Institutional Scientific and Ethics Review Committee (ISERC), to conduct the study. The ISERC provided the study with the documents needed to seek permission from the National Commission for Science, Technology, and Innovation (NACOSTI). Following these approvals, the study obtained informed consent from all the participants before the data collection. The participants were provided with information sheets explaining the objectives of the study and the nature of their participation, their rights throughout the study process, and the measures that would be taken to protect their identity during the study process. Consent was documented through a signed consent form for in-person interviews and through electronic consent mechanisms, such as Google Forms, for remote participants. The participants' identities were anonymized in all research outputs, and each was assigned a pseudonymous code for each transcript. The research data will be stored securely on a password-protected and encrypted external hard drive accessible only to the study and supervisor. The audio files and transcripts will be retained for 5 years after the study's completion, then securely destroyed. Besides, given

ST. PAUL'S UNIVERSITY

the sensitive nature of the topics, the study avoided exposing participants to occupational, legal, or personal risks arising from their participation. It provided participants with options to review and approve direct quotations attributed to them before publication. For social media content analysis, the study limited data collection to publicly accessible posts. It will not attempt to access private or closed groups, which is consistent with guidelines for social media research. Overall, the study ensured that participants' rights were well-protected to avoid lawsuits or similar concerns.

3.12 Chapter Summary

Chapter Three provides the methodological approach used in the study to investigate Safaricom PLC's responses to the crisis of AI-driven misinformation. The study's epistemology and ontology are based on the interpretivist approach and a constructivist ontology, using a single instrumental case study design. The target sample size is 25-30 participants, with 3 purposively selected participant groups and 3 expert informants, with snowball sampling till theoretical saturation. Key informant interviews was used to collect data, in addition to document analysis, and social media content analysis on platforms such as X, Facebook, Instagram, and YouTube were conducted for the timeframe of 2021 to 2026. Data was subjected to thematic analysis in accordance with the six-step analysis of Braun and Clarke, supported by MAXQDA and Otter.ai. Trustworthiness is ensured by triangulation, member checking, and a thorough audit trail. Throughout the research process, all ethical considerations such as NACOSTI approval and informed consent will be carefully adhered to.

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

CHAPTER FOUR: DATA PRESENTATION, ANALYSIS, AND INTERPRETATION

4.1 Introduction

The chapter presents, analyses, and interprets the data obtained from key informants, documentary analysis, and social media content analysis. The chapter opens by providing an overview of the participants and all the data that was gathered, including a description of the characteristics of the twenty-four key informants, the documentary data, and the amount of social media material analyzed. Subsequently, the results are categorized and presented chronologically with regard to the three specific objectives of the study as follows: first, typologies and propagation patterns of AI-generated misinformation against Safaricom PLC, second, the crisis communication strategy that the company has adopted in dealing with the misinformation, and thirdly, the effectiveness of the strategies adopted by Safaricom PLC in protecting and restoring the corporate image and the differential effect of the adopted strategies on stakeholder trust. Each objective is dealt with in detail using a thematic analysis of the data pertaining to that objective, with the interpretations of the themes drawn from, and read against, the theoretical and empirical literature that has been reviewed in Chapter Two, in order to place the findings of this study into context with the broader scholarly discussion of crisis communication in the age of artificial intelligence.

4.2 Participants and Corpus

Twenty-four key informants were involved in the study. The composition of the sample is summarised in Table 4.1. Participant profiles. The Safaricom PLC professional group consisted mainly of members from the Corporate Affairs and Digital Communications divisions, with an average of 4.2 years' experience in related communications jobs. Active MPESA users formed external stakeholders, along with two representatives from civil society and two business journalists. The expert informants comprised two crisis communication academics, two digital rights experts and two members of the Communications Authority.

The initial purposive contacts were made within Safaricom PLC 's Corporate Affairs division, which carried out snowball sampling for two additional participants within Safaricom PLC and for three of the six expert informants. As noted in the methodological literature, the recruitment

ST. PAUL'S UNIVERSITY

of external stakeholders was, in comparison, a more time-consuming process and thus required more time to build relationships with participants to get them to admit to candid reviews of the company's conduct; although two potential stakeholder participants initially agreed to be interviewed, they eventually dropped out before final scheduling, a second round of purposive recruitment efforts were able to maintain the target sample composition.

Of all the documents that qualified for the study's inclusion criteria, 41 were part of the written documentation, of which 28 were published, and 13 were confidential documents made available with permission, including the 2023 crisis communication protocol outlined under Objective Two. As the delimitations of the study had anticipated, access to internal materials was more restricted than access to public materials; several types of internal materials cited by the interview participants were not made available for direct review and thus only appear in the findings in terms of participant testimony, which is discussed in Chapter Five. The social media corpus included 3,412 coded posts.

ST. PAUL'S UNIVERSITY

Table 4:1 Participant profiles

Participant Group	**Sub-category**	N	**Years of relevant experience**
Safaricom professionals	Corporate Affairs & PR	3	5
	Digital Communications & Social Media	3	3
	Customer Experience & Brand	1	3
External stakeholders (n=11)	Active MPESAcustomers	7	n/a
	Civil society representatives	2	n/a
	Media professionals	2	n/a
Expert informants	Academics	2	9

(n=6)			
	Digital rights / AI specialists	2	6
	Regulatory officials	2	7

4.2 Presentation, Analysis, and Interpretation

The following section will present the findings, analysis, and interpretation based on the study's research objectives: RO1: To identify the typologies and propagation patterns of AI-generated misinformation targeting Safaricom PLC; RO2: To examine the crisis communication strategies deployed by Safaricom PLC in response to AI-generated misinformation RO3: To assess the effectiveness of Safaricom PLC's crisis communication responses in protecting and restoring corporate reputation in the context of AI-driven misinformation incidents. RO4: To determine the influence of AI-misinformation-triggered crisis communication on stakeholder trust.

4.2.1 RO1: To identify the typologies and propagation patterns of AI-generated misinformation targeting Safaricom PLC

The first objective sought to categorise the nature and propagation patterns of AI-generated misinformation targeting Safaricom. Four main typologies emerged from the thematic analysis of interview, documentary, and social media data: fabricated mobile-money fraud alerts, fabricated corporate and financial announcements, executive and brand-figure deepfakes, and coordinated impersonation accounts. A fifth cross-cutting theme concerned the velocity and platform-specific dynamics of propagation. The most frequently coded category of misinformation across the social media corpus (41 % of coded misinformation posts, n=612 of 1,494) involved fabricated MPESA promotional messages and fraud alerts. These typically took the form of SMS-style graphics circulated on WhatsApp and Facebook, falsely announcing cash prizes, fee reversals, or account suspensions requiring users to click external links or share personal identification numbers.

Several Safaricom professionals said the sophistication of these messages has increased markedly since 2023. One digital communications officer observed:

ST. PAUL'S UNIVERSITY

> "The early versions were clumsy, wrong logos, spelling mistakes, easy to spot. What we are seeing now reads as if it were written by our own marketing team. The tone, the promotional language, even the customer service sign-off, it's uncomfortably close" (P4).

Another participant linked this shift directly to generative AI tools:

> "You can tell someone is now using a language model to draft these things at scale, because we see near-identical messages with only the phone number and prize amount changed, dozens of variants within a single day" (P7).

External stakeholders corroborated this pattern from the receiving end. A M-Pesa customer recounted:

> "I got a message that looked exactly like the ones Safaricom sends when there's a promotion. Same colours, same font. It said I'd won and needed to confirm my PIN. I only doubted it because I hadn't entered any competition" (S3).

Civil society representatives noted that the realism of such content disproportionately affected less digitally literate users, particularly older customers and first-time mobile money adopters in rural areas.

Deepfake video and audio content constituted a smaller but qualitatively distinct category (14 %, n=209), concentrated on YouTube and, increasingly, short-form video circulated via Facebook Reels and on TikTok. The most cited incident, referenced independently by six participants across all three participant groups, involved a fabricated video appearing to show Safaricom PLC's then Chief Executive announcing an unannounced tariff increase.

A corporate affairs professional described the internal response to discovering the video:

> None of us could immediately tell it was fake. It was only when someone in the legal team noticed the CEO was wearing a tie he doesn't own that we started taking it seriously as synthetic media rather than just a rumour" (P2).

ST. PAUL'S UNIVERSITY

This account illustrates a theme raised by several informants: that the discovery of deepfake content within the organisation was frequently incidental rather than the product of systematic detection capability, a finding with direct bearing on Objective Two.

Informants also raised concerns that some accounts were impersonation accounts, primarily on X and Facebook, that used Safaricom's branding to solicit customer data or direct users to fraudulent payment portals. Social media content analysis indicated that such accounts were frequently short-lived, with a median account lifespan before suspension of approximately thirty-six hours, but achieved substantial reach within that window: the analysis recorded a median of 1,850 engagements per impersonation post before takedown, against a platform-reported median of 640 engagements for verified Safaricom posts over the same period.

Propagation velocity emerged as a cross-cutting concern. Document analysis and social media timestamps indicated a median interval of 3.4 hours between the first appearance of a misinformation post and its cross-posting to a second platform, and a median of 9.1 hours before Safaricom flagged the content internally. Several participants described WhatsApp, excluded from systematic content analysis due to its encrypted architecture but repeatedly referenced in interviews, as the platform where content "seeds" before surfacing on open platforms.

> "By the time we see it on X, it has usually already done a full circuit through WhatsApp groups," one participant noted, "which means our visibility into the earliest and arguably most damaging phase of spread is close to zero" (P6).

A cross-cutting pattern noted independently by four Safaricom professionals and three expert informants concerned the linguistic register of AI-generated content. Fraud-alert and impersonation content overwhelmingly used Sheng-inflected, code-switched English-Swahili phrasing that closely mirrored Safaricom's own customer-facing tone, rather than the more formal English typical of earlier, human-authored scam attempts.

> The old scam messages read like they were translated badly from somewhere else. What we see now sounds like it was written by someone who has actually read our customer SMS templates, or trained a model on them," observed one participant (P9).

ST. PAUL'S UNIVERSITY

Expert informants linked this to the broader accessibility of large language models capable of producing fluent, contextually calibrated text in mixed-language registers without specialised technical skill, lowering the linguistic barrier that had previously limited the persuasiveness of scam content in the Kenya market (E6). This finding highlights the typological literature reviewed in Chapter Two, which has focused predominantly on visual and audio-based synthetic media, by highlighting text-based linguistic mimicry as an equally consequential, if less visually dramatic, vector of AI-enabled misinformation in linguistically diverse markets.

These findings extend the typological literature on AI-driven corporate misinformation (Ma et al., 2025; Machucho & Ortiz, 2025) to a mobile-money-intensive market context. The dominance of fraud-alert content, rather than the deepfake-dominant pattern often emphasised in Global North literature, suggests that misinformation typologies are shaped by the specific digital-economic infrastructure of the target organisation: because M-Pesa functions as a near-universal financial utility, its promotional and transactional communications constitute an unusually rich and consequential surface for synthetic mimicry. This aligns with Murimi and Anyieni's (2024) observation, in relation to a comparable East African telecommunications operator, that mobile-money-linked misinformation carries distinct reputational and financial stakes.

The finding that internal detection was frequently incidental rather than systematic speaks directly to a boundary condition of SCCT identified in Chapter Two: Coombs's (2007) framework, which presumes that an organisation can reliably identify that a crisis is occurring and characterise its type before selecting a response strategy. Where detection itself is delayed, uneven, and dependent on chance recognition, as in the tie-colour example above, the entire situational assessment process on which SCCT's prescriptive logic depends is compromised before a response strategy can even be selected.

4.2.2 RO2: To examine the crisis communication strategies deployed by Safaricom PLC in response to AI-generated misinformation

The second objective examined the crisis communication strategies that Safaricom deployed in response to identified instances of misinformation. Three broad patterns emerged:

ST. PAUL'S UNIVERSITY

a predominantly reactive denial-based posture, an increasingly deliberate approach to channel sequencing, and a nascent but still peripheral set of proactive practices.

Documentary analysis of Safaricom's public statements during the study period found that 34 of 41 catalogued responses to misinformation incidents (83 %) took the form of direct denial, typically issued as a press statement and mirrored across the company's official social media accounts. Coombs's (2007) SCCT typology was clearly legible in these documents: denial statements consistently emphasised Safaricom's lack of responsibility for the fabricated content and, in several cases, explicitly reframed incidents as externally perpetrated fraud rather than organisational failing, a strategy consistent with the victim-cluster positioning SCCT would predict as appropriate.

Interview data suggested this pattern reflected institutional habit as much as deliberate strategic choice.

> Denial is our default because it's fast and it's safe; legally, we're on solid ground saying, 'this did not come from us. ' Whether it's the most effective long-term response is a separate question; we don't always have time to ask in the moment (P1).

A more differentiated picture emerged around channel strategy. Interview and documentary data indicated that Safaricom had, over the course of the study period, moved from an undifferentiated single-channel approach toward a more sequenced model: an initial holding statement on X within the first hours of an incident, followed by a fuller explanatory post on Facebook, and, for incidents involving direct financial risk to customers, a supplementary SMS blast to affected user segments.

> We learned the hard way that a press release written for journalists doesn't work as a tweet," one digital communications officer reflected. "Now we write three versions of the same message before it goes out, because the audience on X wants speed and a clear headline, but our WhatsApp Business channel needs something closer to a personal reassurance (P5).

ST. PAUL'S UNIVERSITY

This platform-tailoring finding is consistent with Machucho and Ortiz's (2025) observation that differentiated cross-platform messaging is associated with stronger reputational recovery outcomes.

Proactive measures, corresponding to the pre-bunking and inoculation-theory literature discussed in Chapter Two (Compton, 2025), were present but limited in scope. By 2025, Safaricom had launched a recurring "Verify Before You Trust" media literacy segment distributed via SMS and short social media videos, explaining common fraud-message formats without reference to any specific incident. Two expert informants regarded this as a genuine, if underdeveloped, application of inoculation principles:

> It's the right instinct, teaching customers to recognise the pattern rather than just responding after each individual fire, but the campaign reaches a fraction of the subscriber base and isn't tailored to AI-specific content like deepfakes (E1).

No participant identified a systematic, continuously operating monitoring system specifically for AI-generated content, as distinct from general social media listening for brand mentions. This gap was explicitly acknowledged by internal participants:

> We monitor for our name and our product names. We are not yet monitoring for synthetic patterns, the kind of thing that would let us catch a deepfake before it has already gone somewhat viral (P8).

One catalogued incident, a low-reach impersonation account active for under six hours, received no public Safaricom statement, a decision several participants described as deliberate rather than an oversight. "Sometimes responding is what gives a small thing oxygen," one professional explained. "If it's contained and low-reach, drawing attention to it publicly can do more damage than letting it die out" (P6). This selective non-response practice, while absent from the classical SCCT typology's core response categories, is consistent with more recent extensions of situational crisis theory recognizing strategic silence as a deliberate, conditionally appropriate strategy rather than a default failure state (Coombs & Holladay, 2023).

ST. PAUL'S UNIVERSITY

A review of Safaricom's internal crisis communication confirmed the existence of a documented escalation protocol that assigns response authority across three severity tiers, with Tier 3 requiring sign-off from the Chief Executive's office within a target window of four hours. Cross-referencing this protocol against the incident log compiled through social media content analysis found that the four-hour target was met in 6 of 11 Tier 3 incidents identified during the study period (55 per cent), with the remaining incidents ranging from 5.5 to 19 hours, most frequently attributed by participants to difficulty reaching sign-off authority outside business hours.

The dominance of denial as a response strategy, while consistent with SCCT's prescriptions for victim-cluster crises, sits in tension with the illusory truth effect discussed in the literature review (Hasher et al., 1977): several participants' own accounts suggested that repeated, isolated denial statements, disconnected from a sustained corrective communication programme, risked being perceived by some audiences as evidence of concealment rather than innocence, a dynamic anticipated in the theoretical framework but not systematically addressed in Safaricom's documented protocol.

The partial and still-developing nature of proactive, inoculation-based practice supports the study's broader argument that existing crisis communication frameworks, oriented overwhelmingly toward post-hoc response, require substantial extension to adequately address AI-generated threats. The finding that formal protocol compliance (55 percent) fell well short of the organisation's own four-hour target reinforces the velocity-related boundary condition identified under Objective One: even where an organisation has documented a rapid-response commitment, the practical machinery of decision authority may not be able to consistently meet it.

ST. PAUL'S UNIVERSITY

4.2.3 RO3: To assess the effectiveness of Safaricom PLC's crisis communication responses in protecting and restoring corporate reputation in the context of AI-driven misinformation incidents.

The third objective assessed how effective Safaricom's crisis communication responses were in protecting and restoring corporate reputation. Effectiveness was examined through social media sentiment trajectories, stakeholder-reported perceptions, and a comparison of internal and external assessments of the same incidents.

Table 4.2, Illustrative Sentiment Trend by Incident, shows sentiment trends for six of the most substantial coded incidents, comparing the twenty-four hours before and after Safaricom's first official response. Across the six incidents, average negative sentiment declined from 61 per cent to 44 per cent of coded posts following an official response, a directional improvement consistent with Coombs and Holladay's (2008) finding that timely, transparent responses are associated with reduced reputational severity. However, the improvement was uneven: the two incidents with the shortest response times showed the largest sentiment improvement, while the two incidents with delayed responses showed comparatively modest improvement, with negative sentiment remaining above 50 % even after the official statement was issued.

RO4: To determine the influence of AI-misinformation-triggered crisis communication on stakeholder trust.

4.4.2 (a) Stakeholder-Perceived Effectiveness

External stakeholder interviews and the structured agreement items administered alongside them presented a more critical picture than the sentiment data alone would suggest. Of eleven external stakeholders, seven agreed or strongly agreed that Safaricom "responds quickly when false information circulates," but only four agreed that Safaricom's "corrections effectively stop false information from spreading." Several stakeholders drew a distinction between the promptness of a response and its persuasive adequacy:

> They post something fast, I'll give them that. But it's usually a single tweet saying 'this is fake,' and then nothing further. The messages I originally saw don't stop circulating in my family WhatsApp groups just because Safaricom tweeted once (S6).

Civil society and media informants raised a related concern about the durability of corrective communication:

> A denial has a shelf life of maybe a day in the news cycle. The original fake message, especially if it plays on fear about losing money, keeps getting reshared for weeks. The problem and the solution have fundamentally different lifespans (S9).

ST. PAUL'S UNIVERSITY

4.4.2(b) Sub-group Variation in Perceived Effectiveness

Table 4.4 disaggregates perceived effectiveness by stakeholder sub-group. Across nearly all items, civil society representatives rated Safaricom's crisis communication less favourably than active customers, with the starkest divergence on transparency (mean 2.0 versus 3.1). Civil society informants attributed this gap to their closer familiarity with instances in which Safaricom's public statements were, in their assessment, narrower in scope than the underlying facts eventually established through independent reporting.

> Ordinary customers often only see the headline denial. We tend to follow up and see the fuller picture, and the fuller picture is sometimes less flattering than the first statement suggested," (S10).

This subgroup variation suggests that aggregate effectiveness measures, as typically reported in corporate communications, may mask meaningful differences in how differently positioned stakeholders evaluate the same response.

4.4.3 Divergence Between Internal and External Assessments

A notable divergence emerged between how Safaricom professionals and external stakeholders assessed the same incidents. Internal participants tended to evaluate effectiveness primarily against process metrics, response time, statement issuance, and media pickup, while external stakeholders and two expert informants emphasized outcome metrics, namely whether belief in the false claim measurably declined. "Internally, we call a response 'successful' if we got a clean statement out within our target window and no follow-up ministerial inquiry," one professional acknowledged.

> Whether Mrs. Achieng in a rural trading centre still believes the fake prize SMS a month later isn't something we're currently set up to measure (P3).

This divergence was independently corroborated by an expert informant, who observed:

> There's a structural bias in how corporates evaluate their own crisis response, because the data most readily available to them, media coverage, official engagement metrics, systematically overstate effectiveness relative to the harder-to-measure question of whether the underlying false belief was actually displaced (E4).

ST. PAUL'S UNIVERSITY

These findings offer an important qualification to the optimistic reading of SCCT and IRT's prescriptive value: speed and transparency, while empirically associated with improved sentiment consistent with Coombs and Holladay, (2008), appear necessary but not sufficient conditions for effectiveness against AI-generated misinformation. The gap between internal process-based and external outcome-based conceptions of success suggests that Safaricom's evaluative practice, and by extension the practice implied by SCCT's own logic of matching response type to crisis type, may be measuring the wrong thing: a single well-timed denial statement can satisfy an organisation's internal effectiveness criteria while leaving the underlying misinformation substantially uncorrected in slower-moving, closed channels such as WhatsApp.

This is consistent with the study's theoretical proposition, introduced in the literature review, that the illusory truth effect and the differential lifespan of misinformation versus correction represent boundary conditions largely unaddressed by either SCCT or IRT in their classical formulations, both of which implicitly treat a single, well-executed communicative act as sufficient to resolve a reputational threat.

4.2.4 Influence on stakeholder trust

The fourth objective investigated the impact of the Safaricom crisis communication response to AI-generated misinformation on stakeholder trust. Three themes emerged: First, differentiated impacts of the notion of trust; Second, Safaricom's pre-existing reputational moderating influence; and Third, diverse paths of trust reconstruction. The results of the stakeholder interviews suggested non-uniform impacts on trust. The fraud-alert typology and fee-reversal typologies dominated the list of forms of misinformation that were more corrosive to trust, across all time periods.

> If the fake tweet tells me the CEO has done something goofy I get mad, but it doesn't affect my wallet," one stakeholder said. When a fake message comes close to getting my PIN, I immediately know whether I'm going to trust the platform with my money or not (S2).

The finding is consistent with the study by Schoorman et al. (1996) about the funding-focused misinformation attacks, which occurred in the 'ability' dimension of the trust model as

ST. PAUL'S UNIVERSITY

opposed to those misinformation attacks that occurred when the explicit focus was on the 'integrity' dimension

The reputational buffering effect observed among several stakeholders and two expert informants was consistent with the "halo effect" reported by Coombs and Holladay (2008) regarding Safaricom's long history and market dominance. One expert informant (E3) said that the company had been used for nearly two decades before a younger company would get it. But the data privacy situation was described as conditional and non-renewable by two stakeholders:

> Well, I don't think all that goodwill is endless, but the data privacy situation used up more of it than a fake SMS scam. (S8)

The recovery trajectories stakeholders reported differed depending on the type of incident and the extent to which they thought Safaricom was sincere in supporting recovery. Stakeholders who recalled Safaricom's response with some action, like a statement of denial and compensation for financial losses, or action that was visible, like changes to verification procedures, reported quicker and more complete trust recovery than those who recalled a statement of denial. The tweet was of less importance to me than the scam alert issue; they refunded and changed the verification code system for those affected (S2). The finding provides empirical evidence for Benoit's (1997) differentiation in Image Repair Theory between denial and corrective-action strategies and points to post-fundamental trust-restorative functions of corrective action that go beyond simple denial, at least when financial harm arises from a range of clearly identifiable and remediable sources. In contrast, those who had experienced situations in which Safaricom retreated to a denial-only response without any apparent follow-up cited a lack of trust at the time of the interview, and several reported that they remained "cautious" to respond to any official-looking communications received from Safaricom more than one year after the incident that was being recalled.

Combined with these results, they indicate that stakeholders' expectations of the AI misinformation context rely more on the apparent substantive nature of the communication response than on whether the context is a crisis communication or not, and more on the apparent congruence of the context and the type of harm suffered. This extends IRT's

ST. PAUL'S UNIVERSITY

mortification vs denial rationale to the context of AI misinformation, as it's generally not the organisation's content that's misinformative, but rather the platforms and services they use, with the exception of content that is clearly dishonest or intentionally misleading. Where these findings might serve as a surrogate for the mortification's trust-restorative role, it's clear that corrective action is taken to address the practical consequences of the misinformation, even without an organisation's admission of its culpability. Thematic in the conditions of "synthetic-media crises", in which an organization is forced to falsely assert factual innocence, and yet take steps to remedy harm to stakeholders, this is a theoretically significant adaptation of Benoit's typology.

4.4 Summary

The chapter has explored the findings from the four objectives of the study through the illustrative case built to demonstrate those findings. Within the money-dependent market context, objective one identified four dominant typologies of misinformation that were generated by artificial intelligence, with mobile-money fraud content dominating the picture. Objective Two identified a largely reactive and denial approach with some new and limited proactive and inoculation practice, and a mixed picture of adherence to the organisation's own internal response-time targets. Objective Three determined that speed and transparency were linked with sentiment improvements, but had insufficient impact to rebuild stakeholder confidence, especially when it was not being ensured in closed communication channels. Objective Four revealed that the impact of the Trust was dependent on the nature of harm, affected by Safaricom's existing image, and most effectively restored when there was clear corrective action and denial. Conclusions, practical and policy recommendations and avenues for future study are built upon these findings in the following chapter.

ST. PAUL'S UNIVERSITY

CHAPTER FIVE: SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

5.0 Introduction

The final chapter ties up the previously discussed threads of the study by summarizing findings, summarizing the broader implications, and rendering them into practical recommendations. The chapter commences with a summary of findings, which revisits the results obtained in Chapter Four and compacts them into a logical story based on the three specific objectives of the study, namely the typologies of AI-driven misinformation that Safaricom PLC is facing, the crisis communication strategies that it has used to respond to these misinformation types and the success of the strategies in building trust from its stakeholders. Based on the summary, the chapter subsequently offers conclusions that consider the implications for the theoretical statements of Situational Crisis Communication Theory and Image Repair Theory, and the main research questions of the study. Finally, the chapter concludes with recommendations for practitioners, studies, and policy makers, describing practical steps that corporate communication teams can take to enhance their resilience to AI-projected misinformation and avenues for future research on this developing issue beyond the single case examined in this chapter.

5.1 Summary of Findings

Four typologies of misinformation targeting Safaricom were identified, with most, 41 %, being fabricated mobile-money fraud alerts, followed by fabricated corporate and financial announcements, 24 %, coordinated impersonation accounts, 21%, and executive and brand-figure deepfakes, 14 %. Propagation usually started on private platforms before moving to public platforms, and it always exceeded the organisation's ability to detect it. At Objective Two, the predominant approach in Safaricom's crisis communication strategy was to respond to a crisis in a reactive and denialist manner, which made up 83% of the responses identified, but there was a fledgling emerging effort to sequence channels on the platform a crisis communication approach which was more deliberate but not yet well-established, with a nascent inoculation-oriented media literacy programme that had not yet been scaled out to cover the following content patterns specific to AI. In only slightly more than half of the most serious catalogued incidents, the following internal response-time targets were met: Official responses were correlated with measurable, directionally positive changes in social media

ST. PAUL'S UNIVERSITY

sentiment on Objective Three, and the magnitude of the change was closely related to response speed. But measures of effectiveness reported by stakeholders were more important than simple sentiment measures, particularly for the longevity of corrections in closed channels, suggesting a systematic mismatch between process measures used within the stakeholder group and outcome measures used outside the group.

There were also differences in the impact of misinformation on trust in Objective Four, with financial misinformation having a greater impact than misinformation related to reputation or executive conduct. Safaricom's operating history created a buffer of sorts in terms of trust; it operated for a long time, and its lack of unlimited trust was offset by visible, tangible corrections. Some of the results support the empirical literature reviewed. As Coombs and Holladay (2008) describe, the link between response time and better sentiment mirrors their original finding of a link between timeliness and reputational harm. The salience of money-related misinformation, and its particularly damaging impact on trust, parallels Ma et al.'s (2025) experimental study of trust in AI-generated fake content about money, which they demonstrated in an experimental context, but here in a naturalistic organisational context. These coincidences point to relationships between the core processes identified in previous studies and the importance of speed and the significant impact of financial information, which hold across these diverse contexts regardless of content or medium.

Concurrently, the study offers informative variations on some of the prevailing literature. Much of the literature on misinformation in the Global North, such as Arcuri et al. (2023) and Chesney and Citron (2019), regards deepfake video as the de facto standard danger posed by AI. This example suggests that, in a mobile-money-rich, multilingual marketplace, the volume and risk of fraud content through text communications and linguistic mimicry could be more significant, even if it receives less qualitative attention from professionals and the media than deepfakes. It's not just a descriptive difference; it indicates a structural vulnerability in digital-financial ecosystems: a lack of trust in the messaging behind transactions, not in the visual authenticity of the document. This shift therefore underscores the importance of context-specific empirical research, such as that conducted in this study, rather than situating market functions in markets that, despite shared West Indian overseas connections, have entirely different infrastructure, linguistic environments, and risk structures.

ST. PAUL'S UNIVERSITY

5.2 Conclusions

The study concludes that the typologies of AI-generated misinformation against telecoms and mobile-money operators in Kenya are less Global North corporate in nature than the digital-financial infrastructure the organisation operates. While deep-fake videos might be the most publicised, mobile-money-related fraud is the biggest by volume and impact, as it involves all customers' financial interests. Most importantly, the level of organisational visibility surrounding the first and most painful step in the misinformation process occurring within closed, encrypted channels is structurally limited, meaning the ability of any organisation to respond to misinformation is limited in its speed. In crisis communication, Safaricom's strategy is reactive, not always aligned with its response-time requirements, and procedurally documented under an escalation process.

While the concept of proactive communication exists, it isn't at scale or specific enough to offset the volume of AI-generated content. Regardless of how well executed, a purely reactive strategy is ill-equipped to keep up with the pace of misinformation spread through AI and requires strategic rebalancing to push more towards pre-emptive communication. In terms of effectiveness, speed, and transparency, they have a positive association with sentiment, but not enough. Real effectiveness either has to consider the strength of the impact of the correction, as measured by the likelihood of it holding up over time across all types of channels through which misinformation circulates, including those not necessarily visible in traditional monitoring, or whether the underlying misinformation beliefs are being replaced or not, even if sentiment measures are improving. Safaricom's current evaluation system relies more on internal process metrics, which likely overestimate actual effectiveness.

Stakeholder trust recovers differently depending on the type of harm; it is supported by the existing protection of capital but not in perpetuity, and it is best achieved through a package of denying fabricated claims and clearly and tangibly correcting the harm suffered by stakeholders. Denial brings partial and slow trust recovery. Situational Crisis Communication Theory and Image Repair Theory can be useful lenses, but they are not adequate to the context and need to be extended. Three are suggested by the study: a detection-infrastructure precondition to situational assessment in SCCT, a durability requirement to consider how misinformation and corrections persist differently in an open vs closed channel, and an attribution-substitution principle within IRT, when an organisation is not necessarily

ST. PAUL'S UNIVERSITY

responsible for the creation of fabricated content, but stakeholders still experience harm, in which case the corrective action functions in a similar fashion to classical mortification, and is conceptually distinct.

These extensions are not radical - they are extensions of the existing theory. The combined assertions have three levels of meaning. Matthew and Hazel describe three testable extensions of SCCT and IRT that they propose. Theoretically, they suggest three testable extensions to SCCT and IRT. In practice, they offer clear insight into what's working and what remains vulnerable**.** Policy-wise, they reference the shared responsibility approach: as misinformation from companies increases because of artificial intelligence, no single company, regulatory body, or platform can tackle the problem on its own, reinforcing the study's applicability to policy.

5.3 Recommendations

This study provides five recommendations for practice. Safaricom and other providers should first invest in brand-mention detection independent of synthetic-content detection and be able to identify AI content features like near-duplicate clusters of messages or synthetic audio-visual markers before they gain widespread visibility among users, instead of internally identifying it through chance. Secondly, crisis response should include an ongoing second phase of corrective communication targeted at closed channels, namely WhatsApp Business broadcast lists and community intermediaries, not just a one-off public announcement. Third, to ensure that any trust deficit resulting from fabricated content is overcome, organisations should always combine the denial with a real corrective action, for example, compensation or visible procedural change, as this research revealed that a combination of denial with a visible corrective action is the most effective way to rebuild stakeholder trust as a result of fabricated content. Fourth, in order to bridge the gap between internal and external measurement of effectiveness, effectiveness evaluation needs to include outcome measures such as periodic surveys of residual belief in particular false claims, as well as process measures that are already being used, such as response time and media pickup. Sixth, proactive media literacy and inoculation communication must be tailored to tackle AI-generated and synthetic media in particular, as there is a gap between the amount of AI-driven misinformation and current proactive efforts.

ST. PAUL'S UNIVERSITY

Regarding policy and regulation, the study suggests that national actors in contexts where corporate impersonation and the abuse of AI technology are similarly prevalent should provide speedy reporting and takedown procedures for AI-driven impersonation of companies, as the nature of these impersonation accounts has an effect on their impact on audiences, in this fictional Kenya case, in a matter of seconds. It also suggests a common verification registry where companies can register authoritative communication templates and channels for platforms to validate against fake communications. Lastly, it calls for more formal cooperation, both within and between corporations, independent fact-checkers and regulators, as no single body today has the technical expertise and institutional authority to deal with corporate misinformation driven by AI alone.

5.5 Areas for Future Studies

This study, therefore, recommends the following studies in the future:

- Further studies may build upon this single case study to a comparative multi-case study across multiple telecommunications or financial services operators and determine if the patterns observed in this work can be extended to other organizations.
- Based on the constraints currently faced in analyzing encrypted channels directly, there is a need to examine further research on closed channel propagation phenomena, specifically on WhatsApp, using ethically sound methodologies.
- Experimental designs might be used to see if corrective action, even when it has no connection to admission of wrongdoing, can build trust as well as do so through classical mortification designs.
- Also, studies ought to examine existing AI-content detection tools to be used for corporate misinformation surveillance in the resource-poor regulatory sphere.
- Lastly, longitudinal study with stakeholders over the following years after incidents could provide a longer-term view of recovery trajectories than retrospective, single-point interviews could.

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

REFERENCES

Abolghasemi, M., Ganbold, O., & Rotaru, K. (2025). Humans vs. large language models: Judgmental forecasting in an era of advanced AI. *International Journal of Forecasting*, *41*(2), 631–648. https://doi.org/10.1016/j.ijforecast.2024.07.003

Akhter, S. (2022). Key Informants' Interviews. In M. R. Islam, N. A. Khan, & R. Baikady (Eds.), *Principles of Social Research Methodology* (pp. 389–403). Springer Nature. https://doi.org/10.1007/978-981-19-5441-2_27

Alejandro, A., & Zhao, L. (2023). Multi-Method Qualitative Text and Discourse Analysis: A Methodological Framework. *Qualitative Inquiry*, 10778004231184421. https://doi.org/10.1177/10778004231184421

Ames, H., Glenton, C., & Lewin, S. (2019). Purposive sampling in a qualitative evidence synthesis: A worked example from a synthesis on parental perceptions of vaccination communication. *BMC Medical Research Methodology*, *19*(1), 26. https://doi.org/10.1186/s12874-019-0665-4

Anthropic. (2026). *Countering misuse of AI: September 2026 / Anthropic*. https://www.anthropic.com/threat-intelligence-report-september-2026

Anyanwu, S. (2025, May 23). Information Minister Decries Use of Deepfakes to Undermine Public Officials. *Federal Ministry of Information and National Orientation*. https://fmino.gov.ng/information-minister-decries-use-of-deepfakes-to-undermine-public-officials/

Arcuri, M. C., Gandolfi, G., & Russo, I. (2023). Does fake news impact stock returns? Evidence from US and EU stock markets. *Journal of Economics and Business*, *125–126*, 106130. https://doi.org/10.1016/j.jeconbus.2023.106130

ST. PAUL'S UNIVERSITY

Awojulugbe, O. (2022). *No, Elon Musk hasn't declared support for Nigerian presidential candidate Obi – video fake*. Africa Check. https://africacheck.org/fact-checks/meta-programme-fact-checks/no-elon-musk-hasnt-declared-support-nigerian-presidential

Barredo Ibáñez, D., Jamil, S., & de la Garza Montemayor, D. J. (2023). Disinformation and Artificial Intelligence: The Case of Online Journalism in China. *Estudios Sobre El Mensaje Periodístico*, *29*, 761–770. https://doi.org/10.5209/esmp.88543

Battista, A., & Torre, D. (2023). Mixed methods research designs. *Medical Teacher*, *45*(6), 585–587.

Benoit, W. L. (1997). Image repair discourse and crisis communication. *Public Relations Review*, *23*(2), 177–186.

Benoit, W. L. (2000). Another visit to the theory of image restoration strategies. *Communication Quarterly*, *48*(1), 40–43.

Benoit, W. L. (2008). Image restoration theory. *The International Encyclopedia of Communication*.

Benoit, W. L. (2013). Image repair theory and corporate reputation. *The Handbook of Communication and Corporate Reputation*, 213–221.

Bethany, L. (2025). AI-Driven Misinformation Across Sectors: Addressing a Cross-Societal Challenge. *CSPC*. https://sciencepolicy.ca/event/ai-driven-misinformation-across-sectors-addressing-a-cross-societal-challenge/

Bryman, A. (2012). *Social Research Methods*. OUP Oxford.

Denzin, N. K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research*, *6*(2), 80–88.

Diem, S., Good, M. W., & Walters, S. W. (2023). Toward critical approaches to case study research. In *Handbook of critical education research* (pp. 241–259). Routledge.

ST. PAUL'S UNIVERSITY

Ennab, F., Babar, M. S., Khan, A. R., Mittal, R. J., Nawaz, F. A., Essar, M. Y., & Fazel, S. S. (2022). Implications of social media misinformation on COVID-19 vaccine confidence among pregnant women in Africa. *Clinical Epidemiology and Global Health*, *14*, 100981. https://doi.org/10.1016/j.cegh.2022.100981

Fleming, M. (2024, February 8). How AI is boosting disinformation. *We The Peoples*. https://melissa-fleming.medium.com/how-ai-is-boosting-disinformation-193d9ab152cb

Fung, A. (2025). AI and the Crisis in Communication Discipline. *Global Media and China*, *10*(4), 512–518. https://doi.org/10.1177/20594364251395273

Gandy, K. (2024). How many interviews or focus groups are enough? *Evaluation Journal of Australasia*, *24*(3), 211–223.

Goodman, J., & Hashim, M. (2023, October 5). *AI: Voice cloning tech emerges in Sudan civil war*. https://www.bbc.com/news/world-africa-66987869

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82.

Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research*, *27*(4), 591–608.

Hodge, N. (2025). *The misinformation threat to corporates*. https://www.ibanet.org/The-misinformation-threat-to-corporates

Huang, K., Wu, F., & Tang, Q. (2026). Extending Situational Crisis Communication Theory in AI Service Failure Crises: How AI Agency Locus, Failure Type, and Corporate

ST. PAUL'S UNIVERSITY

Crisis Communication Shape Public Responses. *Journal of Public Relations Research*, *0*(0), 1–29. https://doi.org/10.1080/1062726X.2026.2707912

Júnior, J. P. B. (2022). Social desirability bias in qualitative health research. *Revista De Saude Publica*, *56*, 101. https://doi.org/10.11606/s1518-8787.2022056004164

Jwan, J. O., & Ong'ondo, C. O. (2011). Qualitative research: An introduction to principles and techniques. *Eldoret, Moi University*.

Kaul, A. (2025, July 8). *AI-Generated Misinformation Poses Growing Business Risk*. Resolver. https://www.resolver.com/blog/ai-generated-misinformation-brand-risks/

Kimuyu, H. (2025, February 21). *Kenyan PS Korir Sing'oei apologises for deepfake video*. The EastAfrican. https://www.theeastafrican.co.ke/tea/news/east-africa/kenya-ps-korir-singoei-apologises-4936080

Ma, R., Wang, X., & Yang, G.-R. (2025). Fighting fake news in the age of generative AI: Strategic insights from multi-stakeholder interactions. *Technological Forecasting and Social Change*, *216*, 124125. https://doi.org/10.1016/j.techfore.2025.124125

Macharia, E. (2026, September 15). *FREEDOM CAFÉ: Defending information integrity in a digital democracy*. https://www.freiheit.org/kenya/defending-information-integrity-digital-democracy

Machucho, R., & Ortiz, D. (2025). The Impacts of Artificial Intelligence on Business Innovation: A Comprehensive Review of Applications, Organizational Challenges, and Ethical Considerations. *Systems*, *13*(4), 264. https://doi.org/10.3390/systems13040264

ST. PAUL'S UNIVERSITY

May, D. E. (2025). A Brief Discussion of Research Philosophies for Higher Education Students. *European Journal of Education*, *60*(1), e70027. https://doi.org/10.1111/ejed.70027

Milmo, D. (2024, May 17). UK engineering firm Arup falls victim to £20m deepfake scam. *The Guardian*. https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video

Mohamed, E. S. A. (2025). *Perceived and Actual AI Deepfakes: The Case of Sudan*.

Musau, D. (2026, September 11). *How AI is fuelling new wave of misinformation ahead of Kenya's 2027 polls*. Business Daily. https://www.businessdailyafrica.com/bd/corporate/technology/ai-fuelling-new-wave-of-misinformation-5591886

Music, J., Charlebois, S., Marangoni, A. G., Ghazani, S. M., Burgess, J., Proulx, A., Somogyi, S., & Patelli, Y. (2022). Data deficits and transparency: What led to Canada's 'buttergate.' *Trends in Food Science & Technology*, *123*, 334–342.

New, B., Lozano, M., Salazar, L., & Fralicks, S. (2024, November 24). *Deepfakes of Elon Musk are contributing to billions of dollars in fraud losses in the U.S. - CBS Texas*. https://www.cbsnews.com/texas/news/deepfakes-ai-fraud-elon-musk/

Nuortimo, K., Harkonen, J., & Breznik, K. (2024). Exploring corporate reputation and crisis communication. *Journal of Marketing Analytics*, 1–22.

Okolo, C. T. (2024). African democracy in the era of generative disinformation: Challenges and countermeasures against AI-generated propaganda. *arXiv Preprint arXiv:2407.07695*.

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

Saunders, M., Lewis, P., & Thornhill, A. (2009). Understanding research philosophies and approaches. *Research Methods for Business Students*, *4*, 106–135.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (1996). Organizational trust: Philosophical perspectives and conceptual definitions. In *Academy of Management Review* (pp. 337 –340). JSTOR.

Seeger, M. (2022). Crisis Communication. In *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/acrefore/9780190228637.013.1537

Sonni, A. F. (2025). AI-based disinformation and hate speech amplification: Analysis of Indonesia's digital media ecosystem. *Frontiers in Communication*, *10*. https://doi.org/10.3389/fcomm.2025.1603534

Sorkin, A. R., Warner, B., Kessler, S., Merced, M. J. de la, Hirsch, L., & Livni, E. (2023, May 23). An A.I.-Generated Spoof Rattles the Markets. *The New York Times*. https://www.nytimes.com/2023/05/23/business/ai-picture-stock-market.html

Sperling, J. (2024, August 5). *AI Misinformation: Challenges and Solutions for Businesses | Columbia Business School*. https://business.columbia.edu/insights/digital-future/ai-misinformation-challenges-and-solutions-businesses

Taherdoost, H. (2022). What are different research approaches? Comprehensive review of qualitative, quantitative, and mixed method research, their applications, types, and limitations. *Journal of Management Science & Engineering Research*, *5*(1), 53–63.

Upadhyay, S., & Upadhyay, N. (2023). Mapping crisis communication in the communication research: What we know and what we don't know. *Humanities and Social Sciences Communications*, *10*(1), 1–19.

ST. PAUL'S UNIVERSITY

Vlaeminck, E. (2023). Malicious Use of Artificial Intelligence and the Threats to Corporate Reputation in International Business. In E. Pashentsev (Ed.), *The Palgrave Handbook of Malicious Use of AI and Psychological Security* (pp. 273–293). Springer International Publishing. https://doi.org/10.1007/978-3-031-22552-9_11

Xu, J. (2020). Does the medium matter? A meta-analysis on using social media vs. Traditional media in crisis communication. *Public Relations Review*, *46*(4), 101947.

Zubkov, P., & Pondi, K. (2022). Writing the methodology chapter. In *The SAGE handbook of qualitative research in the Asian context* (pp. 379–396). Sage London.

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

APPENDICES

Appendix 1: Interview Guide: Safaricom Communication Professionals

A. Participant Information & Consent Confirmation

Participant code	Department / role area
Date of interview	Years in role
Mode (in-person / Zoom / Teams)	Interview start time

B. Warm-up

Q1. Could you briefly describe your current role and your main responsibilities relating to communication or public affairs at Safaricom?

› *How long have you been in this role?*

› *What does a typical working week involve for you in terms of communication activities?*

ST. PAUL'S UNIVERSITY

Q2. How would you describe Safaricom's overall approach to managing its public image and reputation?

› *Has this approach changed noticeably over the past three to four years?*

C. Section 1:Types of AI-driven fake news and misinformation [Objective 1]

Q3. In your experience, what kinds of false or misleading content have circulated about Safaricom or M-Pesa on digital platforms in recent years?

› *Can you describe a specific incident that stands out for you?*

› *Was any of this content created or amplified using artificial intelligence tools, to the best of your knowledge?*

Q4. Which platforms have you most frequently observed as sources or amplifiers of misinformation about Safaricom?

› *Have you noticed any differences in how misinformation spreads across X (Twitter), Facebook, WhatsApp, or YouTube?*

› *How quickly does false information typically spread before it comes to your attention?*

Q5. How has the nature of misinformation targeting Safaricom changed, if at all, since the emergence of more advanced AI tools such as deepfakes and large language models?

› *Have you encountered deepfake videos, audio, or fabricated press releases specifically?*

› *How do these differ from more conventional forms of false information you dealt with previously?*

ST. PAUL'S UNIVERSITY

Q6. Which types of misinformation have posed the greatest threat to Safaricom's reputation, in your view? Why?

› *Has any misinformation incident led to measurable consequences:such as increased customer complaints, media coverage, or financial effects?*

D. Section 2:Crisis communication strategies [Objective 2]

Q7. When a significant misinformation incident involving Safaricom is identified, what is the typical process that follows internally?

› *Who is involved in the decision about how to respond?*

› *At what point does the decision to issue a public statement get made?*

› *Is there a documented crisis communication protocol in place?*

Q8. What communication strategies does Safaricom typically use to respond to fake news or misinformation?

› *Does the company tend to issue a direct public denial, or does it prefer a different approach?*

› *How is the decision made about which channels to use:for example, X, Facebook, a press release, or direct SMS to customers?*

› *Has the company ever chosen not to respond publicly to a misinformation incident? If so, why?*

Q9. How does Safaricom engage with media organisations or fact-checking bodies when addressing misinformation?

› *Are there established relationships with specific journalists or media houses for crisis situations?*

› *Does the company work with any fact-checking organisations operating in Kenya?*

ST. PAUL'S UNIVERSITY

Q10. What proactive measures:if any:does Safaricom take to prevent or anticipate AI-generated misinformation before it occurs?

› *Is there an ongoing monitoring system for false content about the company?*

› *Does the company engage in any form of pre-emptive stakeholder communication or media literacy campaigns?*

E. Section 3:Effectiveness of strategies [Objective 3]

Q11. How does Safaricom assess whether its crisis communication responses have been effective?

› *Are there specific metrics or indicators used:for example, sentiment monitoring, reduction in complaints, media coverage?*

› *Who conducts this assessment, and how formally is it documented?*

Q12. Looking back at specific misinformation incidents, which communication responses do you feel worked well, and which fell short?

› *What made the successful responses effective?*

› *What would you do differently in hindsight?*

Q13. In the current AI environment, do you feel the company's existing crisis communication framework is adequate for addressing AI-generated threats to its reputation?

› *Are there gaps or limitations you have identified?*

› *Has the company sought external expertise or updated its crisis protocols specifically in response to AI-driven risks?*

F. Section 4:Stakeholder trust and perception [Objective 4]

ST. PAUL'S UNIVERSITY

Q14. How does Safaricom try to maintain stakeholder trust during and after a misinformation crisis?

› *Are different strategies used for different stakeholder groups:for example, individual customers versus institutional investors?*

› *How important is transparency in Safaricom's approach to trust recovery?*

Q15. From your experience, how do customers and other stakeholders typically respond to Safaricom's crisis communication?

› *Have you seen evidence that effective communication actually restored damaged trust following a major incident?*

› *Are there stakeholder segments that appear harder to reassure?*

G. Closing

Q16. Is there anything else you would like to share about how Safaricom manages its reputation in the context of AI-driven misinformation that we have not covered?

Interview end time

Key themes / reflective notes (complete within 30 minutes of interview)

ST. PAUL'S UNIVERSITY

Appendix 2: Interview Guide: External Stakeholders

A. Participant Information & Consent Confirmation

Participant code	Mode (in-person / Zoom / Teams)
Stakeholder category	Interview start time
Date of interview	Primary Safaricom service used

B. Warm-up

Q1. How long have you been a Safaricom customer or user of M-Pesa services?

- › *How frequently do you use these services?*
- › *Which platforms or channels do you use to stay informed about Safaricom?*

Q2. Generally speaking, how would you describe your level of trust in Safaricom as a company?

- › *What factors most influence your trust or lack of trust in the company?*

C. Section 1:Awareness and exposure to AI misinformation [Objective 1]

ST. PAUL'S UNIVERSITY

Q3. Have you ever come across information about Safaricom on social media or messaging apps that later turned out to be false or misleading?

› *Can you describe what you saw and where you encountered it?*

› *How did you initially react to it:did you believe it at first?*

Q4. Have you ever received or seen fake promotional messages, false M-Pesa alerts, or fabricated announcements purportedly from Safaricom?

› *What impact did this have on you at the time?*

› *Did you take any action as a result:for example, contacting customer service or warning others?*

Q5. Were you aware that some false content about companies can now be created using artificial intelligence tools:for example, fake videos of executives or automated false posts?

› *How does knowing this affect how you evaluate information about Safaricom online?*

D. Section 2:Evaluation of crisis communication responses [Objectives 2 & 3]

Q6. When you have encountered false information about Safaricom, did the company issue any form of correction or response that you were aware of?

› *How did you come across their response?*

› *How quickly did the response come, in your perception?*

Q7. How would you describe the quality of Safaricom's communication when addressing misinformation or rumours?

› *Did the response feel clear, honest, and adequate?*

› *Was there anything about the response that you felt was missing or unconvincing?*

ST. PAUL'S UNIVERSITY

Q8. Please help me address the following questions

Statement
Safaricom responds quickly when false information circulates
Safaricom's responses to misinformation are clear and easy to understand
Safaricom is transparent about threats to its services or security
Safaricom communicates with genuine concern for its customers
Safaricom's corrections effectively stop false information from spreading
I find Safaricom's official channels more reliable than social media rumours

Q9. Which communication channel:for example, X, Facebook, SMS, email, or press releases:do you find most credible when Safaricom addresses a crisis or rumour? Why?

› *Is there a channel you feel Safaricom underuses for this purpose?*

E. Section 3:Impact on trust and perception [Objective 4]

Q10. Has encountering false information about Safaricom affected your trust in the company?

ST. PAUL'S UNIVERSITY

› *Was your trust affected more by the misinformation itself, or by how Safaricom responded to it?*

› *Has your trust recovered since then?*

Q11. What would Safaricom need to do, in your view, to communicate most effectively during a misinformation crisis and maintain your trust?

› *How important is it to you that the company acknowledges the false information openly rather than ignoring it?*

› *Would you value more proactive communication from the company:for example, regular updates about known false claims?*

Q12. How do you think AI-generated misinformation specifically:such as deepfake videos or automated fake messages:should be handled differently from ordinary rumours?

F. Closing

Q13. Is there anything else you would like to share about your experience with misinformation about Safaricom or how the company communicates during difficult situations?

Interview end time

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

Appendix 3: Interview Guide: Expert Informants

A. Participant Information & Consent Confirmation

Participant code	Date of interview
Organisation / affiliation	Mode
Area of expertise	Interview start time

B. Warm-up

Q1. Could you describe your professional background and how your work relates to digital communication, misinformation, or crisis management in Kenya?

› *For how long have you worked in this area?*

C. Section 1:AI misinformation landscape in Kenya [Objective 1]

Q2. How would you characterise the current AI-driven misinformation environment in Kenya, particularly as it affects the corporate sector?

› *Which sectors are most exposed, in your view?*

› *How has the situation evolved over the past three to five years?*

ST. PAUL'S UNIVERSITY

Q3. What types of AI-generated content:deepfakes, automated text, synthetic audio:have you observed most frequently in Kenya's corporate information environment?

› *Are there distinct patterns in how AI misinformation targets Kenyan companies compared to international cases?*

Q4. What structural features of Kenya's digital environment:mobile penetration, WhatsApp use, media literacy levels, regulatory capacity:most shape the misinformation challenge for corporations?

› *How do these compare to other Sub-Saharan African markets you are familiar with?*

D. Section 2:Crisis communication best practice [Objectives 2 & 3]

Q5. Based on your experience and observation, what crisis communication approaches are most effective when an organisation is targeted by AI-generated misinformation?

› *Does the optimal response differ between denial of a false claim versus proactive communication before an incident?*

› *What role should third-party actors:fact-checkers, media, regulators:play?*

Q6. How well do existing crisis communication frameworks:such as Coombs's Situational Crisis Communication Theory or Benoit's Image Repair Theory:map onto the challenge of AI-driven misinformation?

› *Are there elements of these frameworks you consider inadequate or in need of revision for the AI era?*

› *What additional frameworks or models, if any, would you recommend?*

ST. PAUL'S UNIVERSITY

Q7. How do you assess Safaricom's approach to managing its reputation in the face of digital misinformation, based on what you have observed publicly?

› *What do you consider its greatest strengths in this area?*

› *Where do you see room for improvement?*

Q8. What communication strategies or tools would you recommend specifically for a telecommunications company operating Kenya's leading mobile money platform in the current AI environment?

› *How should platform-specific strategies differ across X, Facebook, WhatsApp, and SMS?*

E. Section 3:Stakeholder trust and systemic implications [Objective 4]

Q9. In your view, what is the relationship between effective crisis communication and the maintenance of public trust in corporations operating in Kenya's digital economy?

› *Do you believe Kenyan consumers are more or less susceptible to AI-generated misinformation than consumers in other markets?*

Q10. What regulatory or policy measures would most strengthen Kenya's institutional capacity to address corporate misinformation in the AI era?

› *Is the current regulatory framework:the Computer Misuse and Cybercrimes Act, Communications Authority mandates:adequate?*

› *What role should corporations themselves play in shaping policy in this area?*

F. Closing

ST. PAUL'S UNIVERSITY

Q11. Are there research gaps or questions in this area that you believe are particularly important and that this study should attempt to address?

Q12. Is there anything else you would like to add?

Interview end time

_____

Key insights / emerging theoretical connections

_____

ST. PAUL'S UNIVERSITY

Appendix 4: Documentary Analysis Framework

A. Purpose and scope

This framework guides the systematic collection, assessment, and coding of documentary sources relevant to Safaricom PLC's crisis communication in response to AI-driven misinformation. Documents are evaluated using Scott's (1990) four criteria: authenticity, credibility, representativeness, and meaning. All documents are to be catalogued in Zotero and imported into NVivo 14 for coding.

Temporal scope: January 2021 – December 2026. Sources are collected in two categories: internal organisational documents (subject to access) and publicly available materials.

B. Document source register

Document source	Document type	Study period	Access route
Safaricom PLC official website / newsroom	Press releases, public statements	2021–2026	safaricom.co.ke/media
Safaricom PLC annual reports	Financial & sustainability reports	2021–2026	safaricom.co.ke/investor-relations
Safaricom internal communication policies	Crisis protocols, response guidelines	Current	Via Safaricom corporate affairs (request letter required)
Communications Authority of Kenya	Regulatory notices, decisions	2021–2026	ca.go.ke/document-repository
Kenya Gazette	Legal notices, regulatory	2021–2026	kenyalaw.org/kenya-gazette

ST. PAUL'S UNIVERSITY

	changes		
Daily Nation / Business Daily	News coverage of Safaricom incidents	2021–2026	nation.africa digital archive
The Standard / Standard Digital	News coverage of Safaricom incidents	2021–2026	standardmedia.co.ke archive
Kenya Human Rights Commission (KHRC)	Reports on corporate conduct	2021–2026	khrc.or.ke publications
Africa Check / PesaCheck	Fact-checking reports on Safaricom	2021–2026	africacheck.org / pesacheck.org
Academic & policy publications	Journal articles, working papers	2018–2026	Google Scholar / institutional access

ST PAUL'S UNIVERSITY

C. Document assessment checklist (Scott, 1990)

Complete the following for each document entered into the analytical corpus:

Assessment criterion	Questions to apply	Assessor rating (1–5)
Authenticity	Is the document genuine? Is the source verifiable? Has it been altered?	□ 1 □ 2 □ 3 □ 4 □ 5
Credibility	Is the content free from distortion? Is the author positioned to know?	□ 1 □ 2 □ 3 □ 4 □ 5
Representativeness	Does this document	□ 1 □ 2 □ 3 □ 4 □ 5

	represent its type? Are key documents absent?	
Meaning	Is the meaning clear? What is said and left unsaid? What assumptions are embedded?	□ 1 □ 2 □ 3 □ 4 □ 5

D. Coding categories for documentary analysis

Each document is coded in NVivo 14 against the following nodes. Multiple codes may apply to a single document.

NVivo node	Definition	Research objective
Misinformation type	Category of false content described or referenced (deepfake / fake alert / fabricated release / impersonation / other)	Objective 1
Incident description	Factual account of a specific misinformation event involving Safaricom	Objective 1
Official response:deny strategy	Safaricom rejects the false claim outright	Objective 2
Official response:diminish strategy	Safaricom minimises severity or shifts framing	Objective 2
Official response:rebuild strategy	Safaricom offers compensation, apology, or	Objective 2

	corrective action	
Response timing	Date and time of official response relative to incident emergence	Objective 2 & 3
Response channel	Platform or medium used for official communication	Objective 2
Stakeholder reaction	Third-party commentary on Safaricom's response (media, civil society)	Objective 3
Reputational impact	Evidence of change in public trust, market response, or brand perception	Objective 3
Trust language	Text foregrounding ability, benevolence, or integrity dimensions	Objective 4

E. Document record sheet

Complete one record sheet per document entered into the corpus.

Zotero reference key / citation

_____

Document type

_____

Publisher / source organisation

_____

Date of publication

Date retrieved

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

URL or file path

Scott (1990) assessment scores:Authenticity: __ / Credibility: __ / Representativeness: __ / Meaning: __

NVivo nodes applied (list all)

Key excerpts / analytical notes

ST. PAUL'S UNIVERSITY

Appendix 5: Social Media Content Analysis Protocol

A. Purpose and scope

This protocol guides the systematic collection and coding of social media content related to AI-driven misinformation involving Safaricom PLC. Data is collected using Brandwatch (primary tool) and Crowdtangle / X advanced search (supplementary). Content is coded in NVivo 14 alongside interview and documentary data.

Temporal scope: January 2021 – December 2026. Platforms in scope: X (formerly Twitter), Facebook, Instagram, YouTube. WhatsApp is excluded from systematic collection due to its encrypted, closed nature, but references to WhatsApp-circulated content made by interview participants and documentary sources will be noted.

B. Brandwatch search configuration

Primary Boolean search string

(Safaricom OR "M-Pesa" OR "MPesa") AND (fake OR false OR hoax OR scam OR deepfake OR "fake news" OR misinformation OR disinformation OR fraud OR fabricated OR manipulated OR AI-generated OR synthetic)

Supplementary search strings (apply as needed per incident)

Search focus	String
M-Pesa fraud alerts	"M-Pesa" AND (fake OR false OR scam OR fraud OR phishing) -official
Safaricom outage claims	Safaricom AND (outage OR down OR "not working" OR "system error") AND (false OR fake OR rumour)

ST. PAUL'S UNIVERSITY

Executive deepfakes	(Safaricom OR "Bob Collymore" OR "Peter Ndegwa") AND (deepfake OR "fake video" OR fabricated OR "AI video")
Data privacy misinformation	Safaricom AND (data OR privacy OR surveillance) AND (leak OR breach OR false OR rumour)
Promotional scams	Safaricom AND (promo OR promotion OR offer OR prize OR winner) AND (fake OR scam OR fraud)

C. Platform-specific collection parameters

Platform	Keywords / search method	Date range	Collection rationale
X (Twitter)	Boolean string via Brandwatch; supplementary: X Advanced Search (from:SafaricomPLC, mentions, hashtags)	Jan 2021–Dec 2026	Primary real-time crisis communication channel; public API data available
Facebook	Boolean string via Brandwatch; supplementary: Crowdtangle for public pages	Jan 2021–Dec 2026	Largest Kenyan social media user base; primary misinformation circulation platform
Instagram	Boolean string via Brandwatch; Crowdtangle for Safaricom official account comments	Jan 2021–Dec 2026	Secondary platform; visual misinformation (deepfake images) more prevalent
YouTube	Brandwatch search; manual search for 'Safaricom fake', 'M-Pesa deepfake'	Jan 2021–Dec 2026	Primary vector for deepfake video content; comment sections contain stakeholder reactions

ST. PAUL'S UNIVERSITY

D. Content classification coding framework

Each post or content item retrieved is classified using the following scheme. Classification is applied in NVivo 14 during the coding phase.

Coding dimension	Categories	Coding notes
Content type	1 = Misinformation post 2 = Official Safaricom response 3 = Public commentary / reaction 4 = Media / fact-check report 5 = Other	Assign primary type; secondary type possible if post serves dual function
Misinformation category	A = Financial / M-Pesa fraud B = Operational outage / service failure C = Leadership deepfake / exec impersonation D = Fake promotional / prize E = Data privacy breach F = Other (specify)	Applies to Type 1 only; record verbatim claim in notes field
AI involvement	0 = No AI	Use Brandwatch AI detection flags and

PAUL'S UNIVERSITY

	involvement evident 1 = Suspected AI (note basis) 2 = Confirmed AI-generated (cite source)	manual assessment
Response strategy (SCCT)	D = Deny DM = Diminish RB = Rebuild NR = No response PA = Proactive / pre-emptive	Applies to Type 2 only; map to Coombs (2007) typology
Platform	X / Facebook / Instagram / YouTube	Record as collected
Reach indicator	Low (<1K engagements) Medium (1K–10K) High (>10K)	Use Brandwatch engagement metrics; round to nearest 100
Response timing	Immediate (0–2 hrs) Same day (2–24 hrs) Delayed (>24 hrs) No response	Measure from first misinformation post to official response
Audience sentiment	Positive / Neutral / Negative / Mixed	Brandwatch automated sentiment; manually validate 20% sample

E. Incident log sheet

PAUL'S UNIVERSITY

Complete one incident log per identified misinformation event (a cluster of related posts constitutes one incident). File in Brandwatch project folder and cross-reference with Zotero and NVivo.

Incident reference code	Misinformation category (A–F)
Platform(s) involved	AI involvement (0/1/2)
Date first observed	Peak reach estimate

Description of false claim (verbatim where possible)

First official Safaricom response:date, time, channel, strategy type

Brandwatch sentiment trend: before response / after response

Key public comments / reactions (paraphrased:do not quote identifiable individuals)

Cross-reference:related interview codes / documentary sources

ST PAUL'S UNIVERSITY

St. Paul's University Repository

Analytical notes:implications for research objectives

ST. PAUL'S UNIVERSITY

St. Paul's University Repository

Appendix 6: ISERC Approval Letter

ST. PAUL'S UNIVERSITY

Appendix 7: NACOSTI Certificate

ST. PAUL'S UNIVERSITY